The London School of Economics and Political Science

Studies in Risk Aversion and Methods in Economics

Svetlana Chekmasova

A thesis submitted to the Department of Economics
for the degree of Doctor of Philosophy

July 2019

Declaration

I certify that the thesis I have presented for examination for the PhD degree of the London School of Economics and Political Science is solely my own work other than where I have clearly indicated that it is the work of others (in which case the extent of any work carried out jointly by me and any other person is clearly identified in it).

I warrant that this authorisation does not, to the best of my belief, infringe the rights of any third party.

I declare that my thesis consists of 24,955 words.

Statement of conjoint work

I confirm that Chapter 2 was jointly co-authored with Ian Walker (Imperial College London), and I contributed 80% of this work.

Statement of inclusion of previous work

I confirm that Chapter 3 was the result of previous study for a Masters of Research award I undertook at the LSE.

Abstract

The papers in this thesis cover a variety of ideas. They are united by the common theme of carefully observing existing relationships in data. In the first chapter, I find that looking at averages can be insufficient. On average, there is not a strong relationship between an individual's ability and her tolerance of risk. However, the joint distribution of the two characteristics shows a great deal of heterogeneity that the average masks. The highest ability individuals are most likely to report middle levels of risk tolerance, whereas those of lower ability are also likely to report extreme values. For those of high ability, therefore, insufficient risk tolerance may prevent them from starting their own business. In the second chapter, I illustrate and quantify the efficiency gain that results from accounting for nonlinearities in the first stage of 2SLS when estimating the second stage parameters of interest. Additionally, I show that in some cases estimating nonlinearly can prevent incorrect inference resulting from misestimation of the underlying first stage data generating process when using a linear method. In the third chapter, I observe and attempt to explain a behavioral puzzle. A standard screening question for rational decision-making screens out a large minority of the sample. Since so many people are affected, it is unlikely to be due to noise or measurement error. I take the answers seriously and develop an explanation.

Contents

List of Figures

List of Tables

Chapter 1

Ability, Risk Tolerance, and Entrepreneurship

1.1 Introduction

Economists have long considered ability to be essential to entrepreneurship. Schumpeter cast the entrepreneur as an innovator, positing that the amount of entrepreneurship observed directly relates to the "quality of the personnel available in a society" (Schumpeter, 1947). More recent work, such as Evans and Jovanovic (1989) and Levine and Rubinstein (2018), emphasizes the impact of intellectual ability on the choice to become an entrepreneur as well as performance once having started one's own business.

We would thus expect that increasing ability would always have a positive impact on entry into entrepreneurship. Indeed, both above-mentioned papers model ability as an input into the entrepreneur's production function and ensure that the probability of choosing that career over salaried employment increases with ability.

However, in two separate cohorts from the National Longitudinal Study of Youth (NLSY), I observe that the marginal effect of ability on the probability of pursuing entrepreneurship is positive only up to a point, and further increases in ability do not result in higher rates of entry. Put differently, a person of average intellect is as likely to be an incorporated business owner as is someone at the high end of the scale. How can we reconcile this observation with the models?

Cognitive skill is not the only personal characteristic that affects career choices. Prior empirical work suggests that a variety of noncognitive personality traits influence who

chooses to become an entrepreneur and their success in this endeavour (de Meza and Southey, 1996; Hurst and Pugsley, 2011). If ability and another trait are jointly determined in the population, as intellectual ability increases across individuals, the other trait would not remain constant. If both characteristics influence the probability of entrepreneurship, it is possible for the total derivative – total impact via direct and indirect channels, which is what we see in raw data – with respect to ability to be zero even while the partial derivative (direct impact of ability *ceteris paribus*) is positive.

I argue that risk tolerance is the relevant trait, accounting for which resolves the discrepancy between models and observation. Levine and Rubinstein (2018) extend their main model to include the direct impact of risk tolerance as well as ability on career choice (the only paper to do so, to the best of my knowledge), modeling each partial derivative as positive. Their approach implicitly assumes the two traits are either independently determined, or at most correlated. If positively correlated, the total derivative would be larger than the partial; if negatively correlated, it would be smaller. Regardless, we would observe a continuous increase (or possibly decrease) across the whole support, not a change from positive slope to zero.[1] To account for such a change we need a change in the correlation between risk tolerance and cognitive skills part of the way through the support – in other words, a nonlinear relationship, which here takes an inverse-U shape.

The main message of this paper is that cognitive skill and risk tolerance are not separable, but instead are jointly determined, and the relationship between them cannot be captured by a simple correlation, as it is nonlinear. Both traits appear relevant for self-selection into entrepreneurship and thus must be considered simultaneously to robustly model and estimate the impact of either. This insight helps to explain part of the gender gap in new business formation, and may be important when designing policies intended to encourage start-up formation.

The rest of the paper proceeds as follows. Section 2 describes the data and necessary background. Section 3 documents the empirical observations. Section 4 describes the Levine and Rubinstein model and extends it to account for the puzzle observed. Section 5 estimates the model and several counterfactual scenarios. Section 6 concludes.

[1] Such a change could be expected if we were nearing P(entrepreneurship)=1, but this is not the case in the data.

1.2 Background

Before proceeding with the analysis, it is important to establish what I mean by "entrepreneur", as this word can be used to describe different kinds of economic activity. For reasons laid out below, the definition used in this paper is an individual self-employed in an incorporated business. It is also necessary to discuss the measure of risk tolerance used in the analysis, as there are multiple potential ways in which it can be measured. I begin by describing the NLSY data.

1.2.1 Data

The two datasets that I use are the National Longitudinal Survey of Youth, 1979 cohort and 1997 cohort (NLSY79 and NLSY97). These are two longitudinal studies following cohorts representative of the US population born in 1957-1964 and in 1980-1984 over multiple years. What is particularly useful about these datasets for my analysis is that aside from data on employment, most participants also took an exam to measure their intellectual ability. Additionally, there is data on some of the participants' noncognitive characteristics.

My measure of intellectual ability is the same as that used by Levine and Rubinstein (2017), the individuals' results on the Armed Services Vocational Aptitude Battery (ASVAB), and in particular their performance on the following subsections: Mathematical Knowledge, Arithmetic Reasoning, Word Knowledge, and Paragraph Comprehension. Together, these subsections are known as the Armed Forces Qualifying Test, or AFQT. A cumulative performance score ranging from 0 to 100 is available for both cohorts. This is a standardized intelligence test developed by the US Armed Forces, which pioneered standardized testing at the beginning of the 20th century (Gallagher, 2003).

The test was administered to participants when they were teenagers, at the beginning of the longitudinal surveys. Data on risk preferences (discussed below) was collected later, once participants had reached adulthood. Both datasets additionally collect information in each wave on employment in the previous year, as well as many other indicators.

1.2.2 What is entrepreneurship?

In this paper, I define an entrepreneur as an individual who reports being self-employed in an incorporated business in the NLSY survey. The reasoning for this is twofold. First of all, the literature seems to have converged toward this definition (discussed below). The second reason is that transformative, innovative companies – those rare enterprises the inception of which governments wish to encourage – are a subset of all incorporated businesses. Arguably, a necessary, though not sufficient, condition for starting such a company is a high degree of mental acuity. If risk aversion constrains high ability people from entering incorporated entrepreneurship, it may also constrain them from entering transformative entrepreneurship, which is surely an even more uncertain pursuit. This could have policy and welfare implications.

It has previously been shown that employees at a start-up tend to have more patent applications than do employees at established firms (Sauermann, 2017). Even though established companies also do engage in R&D, and occasionally may even produce something completely new rather than just marginal improvements, it may be more difficult for established firms to develop and utilize radical rather than incremental innovations due to organizational constraints (Henderson and Clark, 1990). Many large tech companies such as Google expand their knowledge and enter new areas of operation by buying promising start-ups, for example, Deep Mind. As there is some evidence that most of the benefits of technological change are passed on to the rest of society rather than internalized by the individual inventor, their partners, or their financial backers (Åstebro et al., 2014, Nordhaus, 2004), it is in the interest of all of us that the most capable among us innovate and push their ideas into the market.

The innovation and growth aspect of entrepreneurship has been of interest to economists since research in this field began. Schumpeter (1947) first defined the function of the entrepreneur as "...the doing of new things or the doing of things that are already being done in a new way (innovation)." While Knight (1921) did not emphasize invention, he still characterized entrepreneurs as exceptional contributors to economic growth: people who can perceive opportunities more clearly than can others in a world of unpredictable uncertainties, and thus are able to go into business and generate profits. Many theories have cast the entrepreneur as a main driver of economic growth (Baumol, 1990, Murphy et al., 1991, Vandenbussche et al., 2006, Gennaioli et al., 2013).

Empirical verification of the above theories initially simply utilized those who are self-

employed as a proxy for entrepreneurs. However, it emerged that those who are self-employed do not appear to be so different from those who are salaried (Evans and Leighton, 1989; Moskowitz and Vissing-Jørgensen, 2002; Hamilton, 2000). On average, the self-employed do not appear to be exceptional in any observable way – they are not more highly educated, they do not earn more, they are not much younger or older than are the salaried (Levine and Rubinstein, 2017). Additionally, most small businesses do not require very much start-up capital, remain small, and do not aim to grow or become transformative (Hurst and Pugsley, 2011).

Initial attempts to resolve the disconnect between the theorized and observed business owner suggested that perhaps entrepreneurship is not as exceptional an activity as initially supposed, and that the selection criteria into entrepreneurship may be based simply on the appeal of the job to some people. This research emphasizes the potential importance of noncognitive traits in shaping career choices. Work by de Meza and Southey (1996) suggests that those who enter into entrepreneurship are overly optimistic and overconfident, expecting good outcomes more frequently than they get realized, especially when being in control. Hurst and Pugsley (2011) argue that non-pecuniary benefits, such as the opportunity to be one's own boss, play a first-order role in the business formation decision. If such preferences are randomly distributed in a population and are what leads to entry, then certainly the self-employed and the salaried would not look very different. Lazear (2004) proposes a different hypothesis, that entrepreneurs are people who are jacks-of-all-trades with balanced skill sets. In his view, specialists are better off seeking employment, whereas generalists do well as business owners.

Levine and Rubinstein (2017) reconcile the theory with the empirical observations by arguing that not all those who are self-employed should be considered entrepreneurs. They draw the distinction between the unincorporated self-employed and the incorporated "true" entrepreneurs, showing that it is the latter who engage in analytical rather than manual work and "open businesses that are more closely aligned with core conceptions of entrepreneurship than the unincorporated." I therefore utilize this distinction and define entrepreneurs as those who are self-employed in an incorporated business.

1.2.3 Risk preferences

The measure of risk aversion that I use is a self-assessed measure on a 0-10 scale (the 11-point Likert scale), 0 corresponding to most risk averse and 10 corresponding to the most risk tolerant. While this may at first seem questionable, there is prior work showing

that people are good at assessing their own tolerance for risk. Such a measure predicts individuals' performance in incentivized experiments to assess risk aversion, as well as engagement in actual risky activities (Dohmen et al., 2005; Dohmen et al., 2011; Falk et al., 2015). In the NLSY data, risk preferences are elicited in this way at most once per individual, and the measure is taken when participants are adults in the middle of their careers.

The above may be a concern, but there is evidence that personality characteristics, and in particular risk aversion, are (somewhat) stable over the adult life span (Galizzi et al., 2016). Schildberg-Hörisch (2018) provides an overview of the available evidence. The main message that emerges from that work is that, "Individual risk preferences appear to be persistent and moderately stable over time, [although] their degree of stability is too low to be reconciled with the assumption of perfect stability in neoclassical economic theory." In particular, there is some evidence of systematic change in risk preference with age – children tend to be much less risk averse than adults (Levin et al., 2007; Moreira et al., 2010; Paulsen et al., 2011), and after reaching adulthood, as one gets older, one tends to slowly become more risk averse (Sahm, 2012; Dohmen et al., 2017; Josef et al., 2016; Schurer, 2015). However, while the reported *level* of risk tolerance gradually decreases, the *ordering* of individuals within a population remains quite stable. Therefore, although risk preferences are elicited in the NLSY surveys after people have been active in their careers, as opposed to being measured at the onset of adulthood, I believe it is still an informative measure. Since these surveys follow a particular cohort, age effects within each cohort should be minimal and the responses should reflect a true ordering between the different individuals. Partly for this reason, I do not pool the two samples but analyze each separately.

One further consideration regarding the risk measure I use is whether it may be the case that it is not risk tolerant people who enter into entrepreneurship, but it is entrepreneurship itself that makes people less risk averse. This would mean that when I observe people's risk attitudes in the middle or towards the end of their careers, these would be the result of the career choices they had already made, not characteristics that influenced their decisions initially. This seems unlikely to be the case. It is unclear why someone who is initially risk averse would undertake a risky endeavour such as starting their own business in the first place, and later show up as being tolerant of risk. Whether risk tolerance is a factor in who becomes an entrepreneur at all is not definitively settled in the literature. However, studies comparing entrepreneur and non-entrepreneur samples either find that entrepreneurs are on average less risk averse, or no significant difference (Caliendo et al.

(2009), Fairlie (2002), Hvide and Panos (2014), Ahn (2010), Holm et al. (2013)); the literature demonstrates that it is never the case that those who start their own businesses are *more* risk averse than the average population. In addition, it seems strange to suppose that one's level of risk tolerance does not play a role in making the decision of whether to enter entrepreneurship or not, considering that we usually frame decision-making in terms of expected utility. Surely when an individual considers their perceived costs and benefits of being a salaried employee or starting their own business and the likelihood of success, risk attitudes must come into play.

There is one paper which argues that entrepreneurship does make individuals more tolerant of risk, Brachert and Hyll (2014). Using data from the German Socio-Economic Panel (SOEP), the authors show that measures of occupational risk tolerance decrease slightly for the general population between two measurements, in 2004 and 2009, but that they slightly increase for those who become self-employed in the intervening five years. However, the average level of risk tolerance among those who become self-employed is higher than for those who do not (a larger difference than the decrease over time), and slightly higher still for those who were already entrepreneurs in 2004 (for whom there is on average a slight decrease in 2009). In addition, the sample sizes are quite small, so the observed differences may possibly be due to measurement error or in the selection of the particular sample. Overall, it appears that risk tolerance is likely to be a factor in who chooses to become an entrepreneur, and not that entry into self-employment systematically affects an individual's level of risk aversion.

1.3 Ability, risk tolerance, and entrepreneurship

The main observation of this paper is that highly able individuals may be constrained from entering into incorporated entrepreneurship by their lack of sufficient risk tolerance. Table 1.1 shows that on average, those of high ability (AFQT > 50) do not appear to be at a disadvantage with regards to other noncognitive characteristics. They seem to have better mental health, feel more in control of their lives, have higher self-esteem, and are less likely to be shy. As described below, risk tolerance thus seems a more likely candidate for an occupational choice friction.

Figure 1.1 shows the fraction of person-year observations in which people of each AFQT category have been self-employed in an incorporated business, or in any business (including unincorporated). This shows the importance of distinguishing between entrepreneurs and

Noncognitive characteristics by ability

Characteristic	Scale	Low ability	High ability difference
NLSY79			
Rotter locus of control	4-16, lower=feel more in control	9.174539	-1.395348***
Rosenberg score	6-30, higher=higher self-esteem	21.54127	2.269158***
Sociability at age 6	1-4, low=shy	2.257194	.1897865***
Sociability as adult	1-4, low=shy	2.862089	.0625491***
Pearling mastery of own fate	7-28, higher=greater mastery	21.63789	1.284622***
TIPI extroverted, enthusiastic	1=disagree to 7=agree	5.079556	-.1059296**
TIPI critical, quarrelsome	1=disagree to 7=agree	3.309895	-.3508103***
TIPI dependable, self-disciplined	1=disagree to 7=agree	6.002038	-.0243869
TIPI anxious, easily upset	1=disagree to 7=agree	3.408546	-.5287217***
TIPI complex, open to new experiences	1=disagree to 7=agree	5.267318	-.0633704
TIPI reserved, quiet	1=disagree to 7=agree	4.413255	-.6801557***
TIPI sympathetic, warm	1=disagree to 7=agree	5.597597	-.1136268***
TIPI disorganized, careless	1=disagree to 7=agree	2.624067	-.2538915***
TIPI calm, stable	1=disagree to 7=agree	5.432004	-.0385826
TIPI conventional, uncreative	1=disagree to 7=agree	3.383318	-.1836707***
NLSY97			
Mental health	5-20, lower=more emotional problems	13.42447	.5706225***
TIPI extroverted, enthusiastic	1=disagree to 7=agree	3.641159	.1680063*
TIPI critical, quarrelsome	1=disagree to 7=agree	2.140416	.0916625
TIPI dependable, self-disciplined	1=disagree to 7=agree	4.425706	.0548179
TIPI anxious, easily upset	1=disagree to 7=agree	2.464091	-.4791484***
TIPI complex, open to new experiences	1=disagree to 7=agree	4.078504	.1123305
TIPI reserved, quiet	1=disagree to 7=agree	2.736751	-.3066363***
TIPI sympathetic, warm	1=disagree to 7=agree	3.836305	.222942**
TIPI disorganized, careless	1=disagree to 7=agree	1.527489	.1893688**
TIPI calm, stable	1=disagree to 7=agree	3.85686	.1441222
TIPI conventional, uncreative	1=disagree to 7=agree	1.657751	-.0819739

Table 1.1: Differences in noncognitive characteristics between higher and lower ability individuals. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

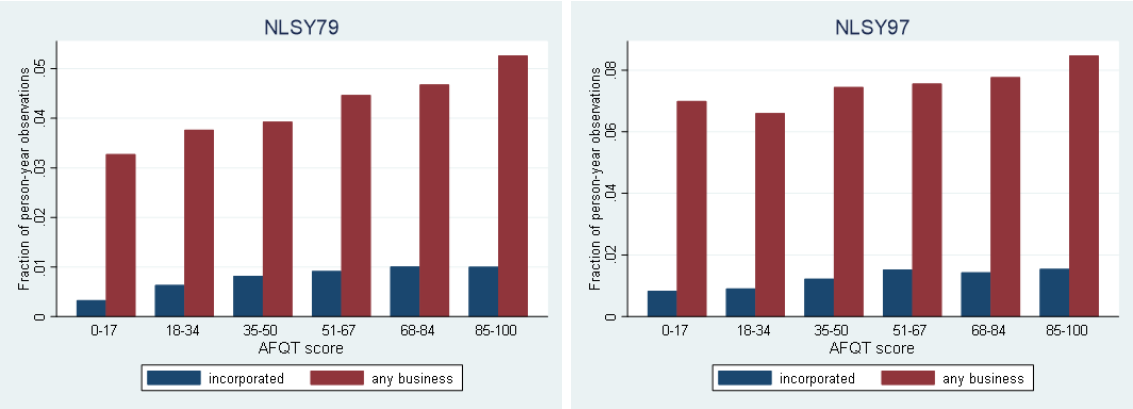


Figure 1.1: Likelihood of entrepreneurship by AFQT score

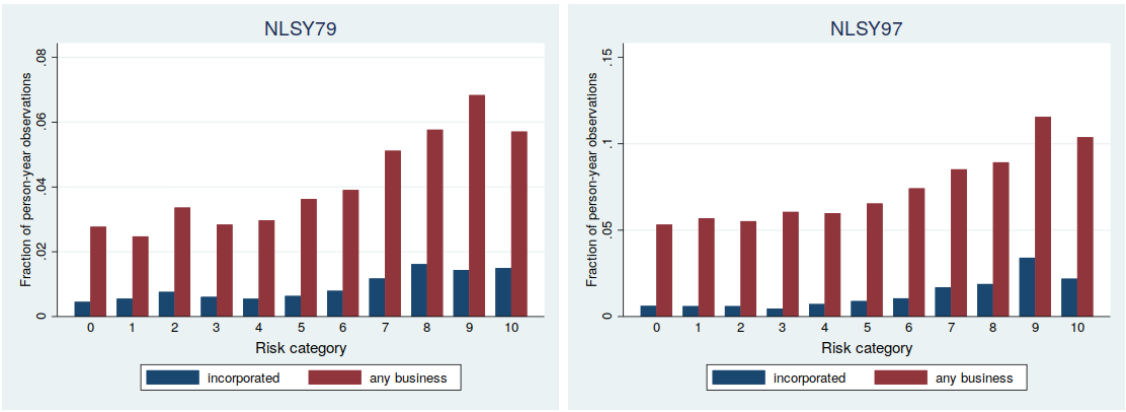


Figure 1.2: Likelihood of entrepreneurship by risk tolerance

other self-employed. We see that while higher intellectual ability is associated with greater entry into self-employment in general, its marginal impact on the likelihood of being an incorporated entrepreneur is not stable. In particular, the marginal effect of ability is positive for those who score 50 or below – the fraction of person-years increases with AFQT score – but it is statistically indistinguishable from zero for those who score above 50 (shown in Table 1.2). Someone who scores a perfect score on the AFQT thus seems as likely to become an incorporated entrepreneur as someone who only scores half the marks.

A useful question to ask here is what benchmark shape we should expect to see. If increasing cognitive skills increases the probability of entrepreneurship, and if ability is determined independently of any other characteristic, the shape should trace out part of a CDF-like curve[2]. Only part because even the highest levels of ability in the sample are insufficient to guarantee pursuing entrepreneurship; the highest probability is still lower than even 5%. We would thus expect to see a convex, increasing shape. Instead, we observe a concave shape, indicating that one or both of the above-mentioned assumptions must not hold.

[2]Further discussion in Section 1.4

Figure 1.2 shows that risk tolerance is more in line with the benchmark. The fraction of person-year observations where an individual is engaged in running an incorporated business remains flat for low levels of risk tolerance, and additional risk tolerance only increases likelihood of entrepreneurship at higher levels. This is also true for self-employment in general. Additionally, from figure 1.3 we can see that a greater proportion of the NLSY97 cohort consider themselves relatively risk tolerant than does the NLSY79 cohort, and members of that cohort do appear to engage in self-employment (incorporated and not) to a greater extent.

If risk tolerance positively influences entry into entrepreneurship, and if it decreases at higher levels of AFQT, this would explain why the probability of entrepreneurship dips to a flat shape as AFQT increases. At the same time, if AFQT positively influences the probability someone becomes an entrepreneur, and it is lower at high levels of risk tolerance, we would expect that to "pull down" the risk tolerance curve as well. Because this curve appears to be "pulled down" less, this may indicate that risk tolerance has a larger direct impact on who becomes an entrepreneur than does cognitive skill.

Table 1.2 shows estimates of the slopes illustrated in figures 1.1 and 1.2. The correlation between higher ability (AFQT) and likelihood of entering into incorporated entrepreneurship is positive and statistically significant for low values of AFQT, but the estimated coefficients are an order of magnitude lower and not statistically significant for high values of AFQT – a concave shape. The point estimate is actually negative for the NLSY97 cohort. The opposite is true for risk tolerance. At low values, a marginal increase in risk tolerance is not associated with any higher participation in incorporated entrepreneurship, but there is a positive and statistically significant effect for those whose value of risk tolerance is 5 or higher. To get a sense of the magnitude of the estimates, the mean value of the incorporated variable (fraction of person-years incorporated) is 0.00826 in the NLSY79 and 0.01225 in the NLSY97. The slope on low values of AFQT can thus be interpreted as an increase of 17% of the average value for a 10 point increase in AFQT in the NLSY79, and an increase of 9.5% of the average value in the NLSY97.

To see why too-low risk tolerance likely hinders the mostly highly able individuals from pursuing a career as an entrepreneur, we need to look at the joint distribution of the two characteristics. This is shown for both samples in figure 1.3. The nonlinear nature of the relationship becomes apparent and is summarized in figure 1.4, an inverse-U shape. Essentially, the variance of risk tolerance decreases with ability. For both cohorts, the distribution is trimodal at low levels of AFQT, but becomes unimodal as AFQT increases.

Marginal effects associated with entry		
NLSY79	Low values	High values
AFQT	.000142***	.0000392
	(.0000307)	(.0000532)
Risk tolerance	.0003234	.0020571***
	(.0003641)	(.0003591)
NLSY97	Low values	High values
AFQT	.0001165**	-.0000172
	(.0000545)	(.0000859)
Risk tolerance	.000122	.0031404***
	(.000599)	(.0004889)

Table 1.2: Results showing the relationships described in figures 1.1 and 1.2. The table shows the estimated marginal effects of risk tolerance and ability on entry into incorporated entrepreneurship estimated using simple OLS with a single regressor (plus constant) and standard errors clustered at the individual level. 'Low values' indicates AFQT ≤ 50 or risk tolerance < 5, and 'high values' indicates AFQT > 50 or risk tolerance ≥ 5. Standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

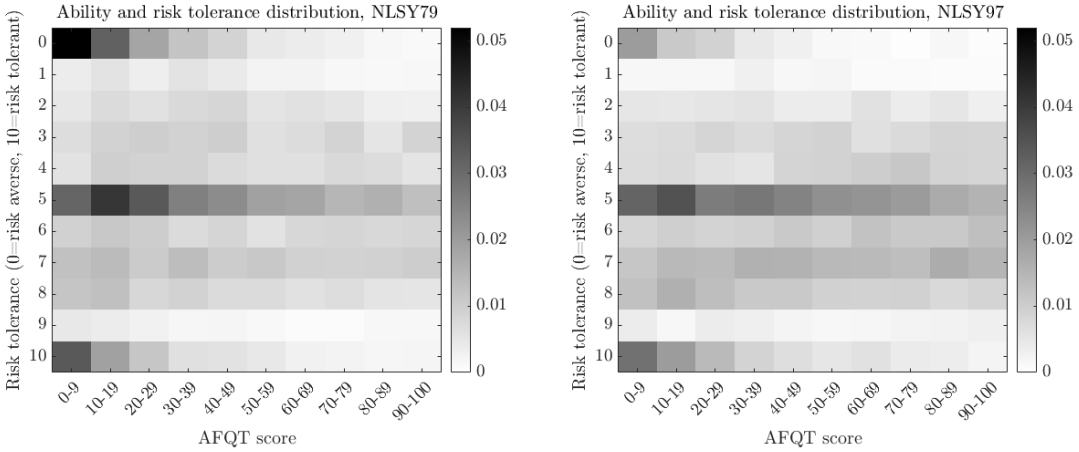


Figure 1.3: Distributions of ability and risk preferences in both samples

As ability increases, relatively fewer and fewer individuals can be found at either extreme of risk preferences. Additionally, there are fewer individuals who score highly on the AFQT than there are those whose score is lower. Not only is the distribution of ability skewed, but few of the highly able are highly risk tolerant. Thus risk tolerance is likely to have particularly strong repercussions for how many high-skill entrepreneurs there are.

In both of the data sets, those who score below 20 on the AFQT are much more likely to report a 0 or a 10 level of risk tolerance than those who score above 80. The pattern generally holds across the ability distribution. In the NLSY79, the proportion of people who report a risk tolerance value of 8 or above (so, those individuals who are least risk averse) among those who score under 20 on the AFQT is .25. Among those scoring 21-40 this is .17, 41-60 it is .16, 61-80 it is .14, and above 80 the fraction is .15. In each successive group there are fewer and fewer individuals, as more people have lower scores than have

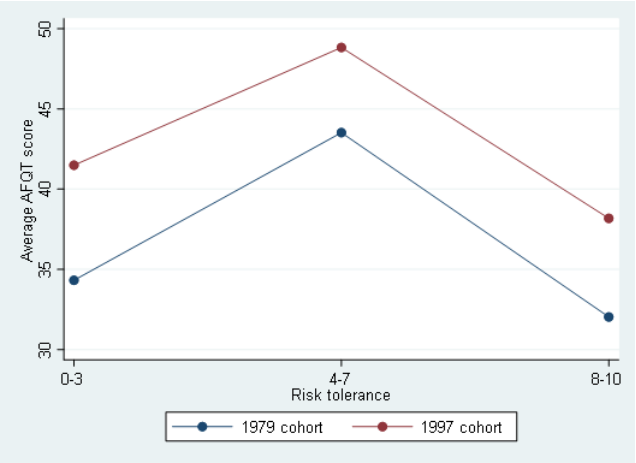


Figure 1.4: Average AFQT score at different levels of risk tolerance

higher scores. For the NLSY97, the corresponding proportions are .31, .25, .19, .19, and .17. Again, each successive group is smaller than the previous one.

Researchers have previously investigated the relationship between intelligence and risk tolerance (Dohmen et al., 2010; Oechssler et al., 2009; Frederick, 2005; Benjamin et al., 2013), finding some evidence of a positive correlation. In the NLSY79, it is true that there is overall a slight but statistically significant positive correlation. In the NLSY97, the correlation is very close to 0 and not statistically significant. What the literature has not considered, to the best of my knowledge, is that the relationship is actually nonlinear, as can be seen in figure 1.4 – the correlation is positive at low levels of risk tolerance and becomes negative once risk tolerance increases beyond 5. It is also clear from figure 1.3 that more important than the change in average risk preferences is the striking heterogeneity of preferences at different levels of intellectual ability. It is not that more and less able people differ so much in their risk tolerance on average, but the less able people tend more towards the extremes, whereas those of higher ability tend to report middle values. Of course, it is possible that this pattern comes about precisely because those of lower ability over-simplify or those of higher ability over-analyze, but I see no reason why they would do so in their survey answers and not in the way they approach choices in their lives. The fact that this heterogeneity is clearly observable in the data for both cohorts indicates that this may be a fundamental quality of how human attitudes toward risk vary with intelligence, and bears further investigation in future research.

Table 1.3 demonstrates the statistical significance of the inverse-U shape shown in Figure 1.4. It is possible to model the relationship between risk tolerance and ability quadratically or using some other functional form. However, the main feature for which it is necessary to account in order to use both as regressors in a linear model that produces meaningful

coefficient estimates is the change in correlation from positive to negative along the risk tolerance support. The simplest way of doing this is to allow for the correlation between the two features to be different when risk tolerance is low and when it is high. I use 5 as the cut-off value, since that is close to the average value of risk tolerance at all levels of AFQT.

Relationship between risk tolerance and ability

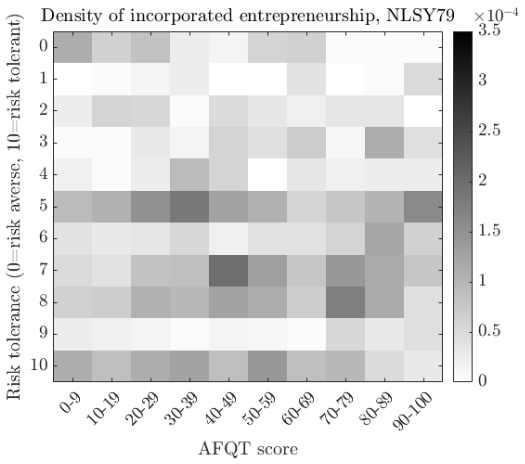
NLSY79	Unconditional		Cond. (risk tol. > 5)		Cond. (risk tol. ≤ 5)	
AFQT	.0059464***	.0059808***	-.0152987***	-.0113716***	.0170295***	.0125196***
	(.0011878)	(.0016345)	(.000872)	(.0012718)	(.0002001)	(.0002760)
Controls	No	Yes	No	Yes	No	Yes
NLSY97	Unconditional		Cond. (risk tol. > 5)		Cond. (risk tol. ≤ 5)	
AFQT	-.0009333	.0000515	-.0137781***	-.0102702***	.008745***	.0057604***
	(.0011061)	(.0015296)	(.0008294)	(.0011972)	(.0002407)	(.0003348)
Controls	No	Yes	No	Yes	No	Yes

Table 1.3: Individual-level regression of risk tolerance on AFQT score. Note that the conditional results are not sensitive to choice of cut-off value. Controls included are: the respondent's family's income in childhood, whether it was a two-parent household, mother's and father's education, and race and sex dummies. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

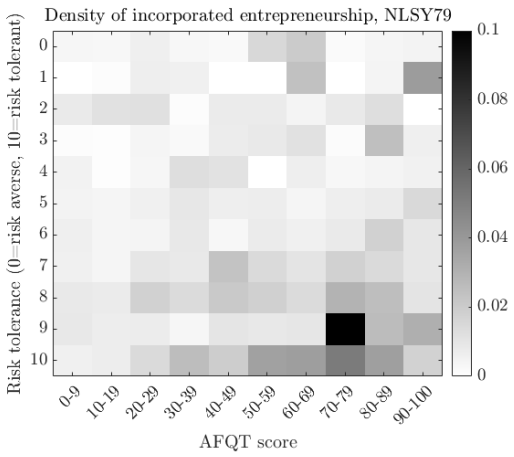
The first two columns of Table 1.3 show the average change in risk tolerance with AFQT, with and without controlling for other characteristics. The controls included are: the family income when the respondent was a child[3], whether the respondent grew up in a two-parent household, the mother's and father's education, and dummies for race and sex. While there is an overall small but statistically significant increase in risk tolerance with AFQT in the NLSY79 data, this is not the case in the NLSY97. These unconditional estimates mask the important nonlinearity present in the data, averaging together highly statistically significant positive and negative slopes. These can be seen in the last four columns of Table 1.3.

To put everything together, figure 1.5 displays where entrepreneurs can be found in two different ways. Subfigures 1.5(a) and 1.5(c) display the total number of person-years in which people of a particular level of AFQT and risk tolerance report running their own incorporated business divided by the total number of person-years in the sample. We can immediately see how rare entrepreneurship is in both samples. We also see that there are very few entrepreneurs below a particular level of risk – it seems to be a phenomenon primarily found in the population of those who consider themselves to be a 5 or higher on the risk scale. This pattern is particularly striking in the NLSY97 distribution, but can also be seen in the NLSY79.

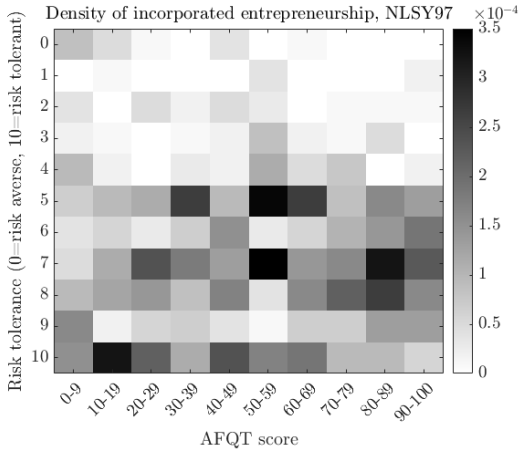
[3]I take the family income data reported in 1979 or 1997, depending on the data set, and if it is missing, I take the data from one or two years later, all converted to 2010 dollars.



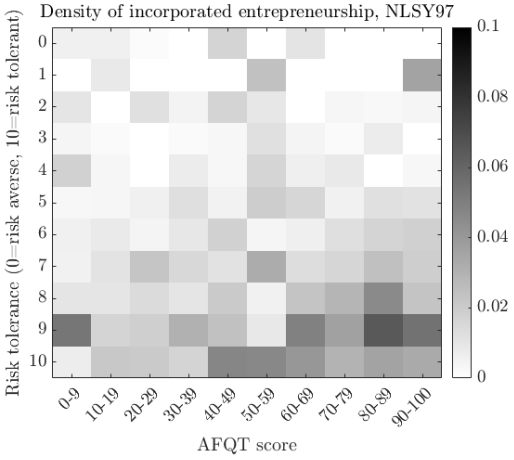
(a) Absolute distribution, NLSY79



(b) Relative distribution, NLSY79



(c) Absolute distribution, NLSY97



(d) Relative distribution, NLSY97

Figure 1.5: Distribution of incorporated entrepreneurship in both samples. Figures labeled 'absolute' show the number of entrepreneur person-year observations in each box as a fraction of the entire sample. Figures labeled 'relative' show the proportion of person-years as a fraction of the total person-years that fall in that box.

Subfigures 1.5(b) and 1.5(d) account for the fact that there is a different total number of people with each pair of values (within each box) and show where entrepreneurs are relatively over- and underrepresented. The values displayed are the number of incorporated person-years in a given box as a fraction of total person-years in that box[4]. These subfigures give some idea of the surface we would like to estimate – the expected probability of entrepreneurship for an individual with a given level of risk tolerance and cognitive ability.

[4]The number of person-year observations in each box differs partly because some combinations are more prevalent than others and partly because individuals may be missing some years of data.

1.4 Model

The model proposed by Levine and Rubinstein (2018) (henceforth LR) can help us to estimate the surface of interest. I begin by summarizing the model as presented in the paper and then offer an extension to reason about the observations presented in the previous section.

1.4.1 LR model summary

LR model the choice to enter into (incorporated) entrepreneurship, salaried employment, or (unincorporated) self-employment. They show that under risk neutrality, entrepreneurs are selected positively on cognitive ability, whereas self-employed individuals are selected negatively. LR offer an extension of the model that accounts for risk aversion and show that including this does not qualitatively alter the self-selection into careers. The extension only changes the levels of the relevant ability cut-offs, keeping the ordering the same. LR thus focus on the insights that they can generate using the risk-neutral model. This paper argues that the way the ability cut-off changes with risk tolerance interlinks with the joint distribution of both characteristics in an important way, so I focus on LR's extended model – in particular, on the choice between entrepreneurship and salaried employment, as in the LR model it is the lowest-ability individuals who select into self-employment, and that choice is unaffected by risk preferences.

LR define an individual's constant absolute risk aversion utility function in the following way:

$$V_{Ji} = -\exp\left(-\tau_i \cdot I_{Ji} \cdot \exp\left(\delta_{Ji}\right)\right), \qquad J \in \{E, S, U\} \tag{1.1}$$

where the letters correspond to being an entrepreneur, salaried, or self-employed, respectively. I_{Ji} is the individual's income. The parameter τ_i describes the degree of risk aversion, taking positive values for risk averse individuals and decreasing towards 0 as the utility function nears risk neutrality. The parameter δ_{Ji} describes the nonpecuniary benefits of each career choice; LR set $\delta_{Ei} = 0$, so the values are relative to entrepreneurship.

In entrepreneurship, the production function is:

$$Y_i = \theta_i K_i^{\alpha}(1 + \nu_i), \qquad \nu_i \sim N\left(0, \frac{\sigma^2}{\theta_i K_i^{\alpha}}\right) \tag{1.2}$$

The expression for the variance of the productivity shock comes from the assumption that

the variance-to-mean ratio of output is constant, so that splitting a firm does not affect the variance of the aggregate output. In the above, θ_i denotes an individual's entrepreneurial ability, and K_i is the amount of capital the individual chooses to invest into production.

Income I_{Ei} is modeled in the standard way as $Y_i - rK_i$, where r is an interest rate higher than 1. By plugging these expressions into the utility function and taking the expected value, we get the following:

$$\mathbb{E}\{V_{Ei}\} = -\exp(-\tau_i \cdot \{\underbrace{\left(1 - \frac{\tau_i \sigma^2}{2}\right)}_{\gamma_i} \cdot \theta_i K_i^{\alpha} - rK_i\}) \tag{1.3}$$

Maximizing expected utility, the entrepreneur sets the optimal level of capital:

$$K_i^* = \left(\frac{\alpha \theta_i \gamma_i}{r}\right)^{\frac{1}{1-\alpha}} \tag{1.4}$$

γ is a convenient shorthand summarizing the combined impact of both risk preferences (τ) and the risk inherent in production (σ^2); it increases in risk tolerance and decreases in risk. Note that in the model, the value of γ is constrained to be between 0 and 1 for entrepreneurs. Indifference between salaried employment and entrepreneurship occurs when the expected utility of entrepreneurship, evaluated at K^*, is equal to the nonstochastic utility of salaried employment. The income from salaried employment is equal to one's effective human capital, $\theta_i^{\rho_J} \cdot e^{\epsilon_{Ji}}$. Here ϵ_{Ji} represents employment-specific skills and is 0 for entrepreneurship while being (on average) positive for other types of employment. It is assumed to be uncorrelated with ability. LR allow "abilities useful in entrepreneurship to also be productive in salaried employment" and assume $0 \leq \rho_U < \rho_S \leq 1$. $\mathbb{E}\{V_{Ei}\} = V_{Si}$ when the following holds true:

$$\begin{aligned} &-\exp(-\tau_i \cdot \{\gamma_i \cdot \theta_i K_i^{*\alpha} - rK_i^*\}) = -\exp(-\tau_i \cdot \{e^{\delta_{Si}} \cdot \theta_i^{\rho_S} \cdot e^{\epsilon_{Si}}\}) \quad \Rightarrow \\ &\Rightarrow \quad \theta^{\rho_S - \frac{1}{1-\alpha}} = \left(\frac{\alpha}{r}\right)^{\frac{\alpha}{1-\alpha}} (1-\alpha) \cdot \gamma^{\frac{1}{1-\alpha}} \cdot \exp(-\delta_{Si} - \epsilon_{Si}) \end{aligned} \tag{1.5}$$

When the left-hand side is smaller than the right-hand side, an individual chooses to enter into incorporated entrepreneurship. Otherwise, they will choose salaried employment (assuming their ability is not below the cut-off at which they choose self-employment, $\theta_i < \exp\{\left[(\delta_{Ui} - \delta_{Si}) + (\epsilon_{Ui} - \epsilon_{Si})\right]/\rho_S\}$). Figure 1.6 shows what the decision looks like for any individuals who share the same values of α, ρ_S, r, δ_{Si}, and ϵ_{Si}. If the person's values of θ and γ are in the shaded region, it is optimal for that person to become an entrepreneur; in the white area, they maximize expected utility by choosing salaried employment (or self-employment at low levels of ability). The figure also shows that in the corresponding

log-log plot, the relationship is linear rather than curved; this is a useful observation for later estimation of the model.

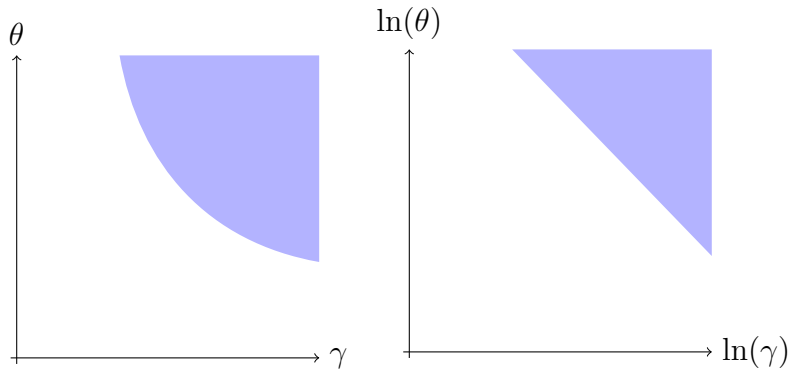


Figure 1.6: The shape of the blue shaded area shows the range of ability and risk tolerance in which an individual with particular parameter values would choose to enter into incorporated entrepreneurship in the LR model.

1.4.2 Digging deeper

The LR model, as presented, gives us a cut-off value based on which an individual will make her career decision. This is informative, but it does not let us reason about the change in the probability of entrepreneurship as ability or risk tolerance increase (Figures 1.1 and 1.2). The model described in the previous section predicts that below a certain point, the probability is zero, and beyond that point, the probability jumps to 1. This does not answer how we should expect the gradient to behave with respect to the two features of interest, either the partial derivatives or the total derivatives. An understanding of the shape of the surface that generates the indifference cut-off is necessary to connect the model to the empirical observations.

We can use the LR model to think about how the probability of entrepreneurship responds to increasing ability or increasing risk tolerance by keeping track of the stochastic element in the model. The resulting probability expression describes a shape for the surface shown in Subfigures 1.5(b) and 1.5(d) – assuming that the other parameters are determined independently of ability and risk tolerance. Under that assumption, different individuals can be valid counterfactuals for each other in expectation, differing only on the quantities of interest. I focus my analysis on ability and risk tolerance, but the same approach can be used to reason about the effects of relaxing this assumption and the impact of other parameter changes.

I begin by deriving the expression for $P(V_E > V_S)$, the probability that the utility one

realizes in entrepreneurship is higher than that obtainable in salaried employment. We can simplify this problem by noting this is equivalent to $P(I_E > e^{\delta_S} \cdot I_S)$, as the utility is a strictly increasing transformation of those two values. To get I_E, the income of the entrepreneur, we need the amount she produces. The would-be entrepreneur sets her optimal level of capital as before, based on expected utility; if she puts that capital into production, she will actually produce $\theta K^{*\alpha}(1+\nu)$, the value of which depends on the realization of ν. I_E is thus a stochastic object, a linear transformation of the normally-distributed ν.

Plugging in the expressions for income in each occupation gives the following:

$$P(I_E > e^{\delta_S} \cdot I_S) = P\left(\left(\frac{\alpha\gamma}{r}\right)^{\frac{\alpha}{1-\alpha}} \theta^{\frac{1}{1-\alpha}}(1+\nu-\alpha\gamma) > e^{\delta_S} \cdot \theta^{\rho_s} \cdot e^{\epsilon_S}\right) \tag{1.6}$$

This expression can now be rearranged with the stochastic ν on the left and all the other terms on the right:

$$P\left(\nu > \underbrace{e^{\delta_S+\epsilon_S}\left(\frac{\alpha\gamma}{r}\right)^{\frac{-\alpha}{1-\alpha}} \theta^{\rho_S-\frac{1}{1-\alpha}} + \alpha\gamma - 1}_{\text{RHS}}\right) \tag{1.7}$$

The above expression describes a surface, a normal CDF on a multivariate support. It gives the probability that realized utility in entrepreneurship is higher than salaried utility for each combination of parameter values. The point at which an individual is indifferent between the two occupations depends on that person's level of risk tolerance. A risk-neutral person would be indifferent when the probability is .5. Someone more risk-averse would need a better-than-even chance that V_E turns out higher and would prefer the nonstochastic salaried option if the odds were 50-50. When the RHS is set equal to $\gamma - 1$, we get back the indifference expression in equation 1.5. Note that $\gamma \in (0,1)$ and increases in risk tolerance. When $\gamma \to 1$, the RHS $\to 0$, so $P(V_E > V_S) \approx .5$. As risk tolerance decreases towards zero[5], this probability needs to increase in order for the person to remain indifferent.

The variance of ν depends on ability and on the capital used in production, making it difficult to reason about the ability gradient. To account for this, we can transform the left-hand side to produce a standard normal. Let $z \equiv \sqrt{\theta K^{*\alpha}}\nu$ such that $z \sim \mathcal{N}(0,1)$.

[5]Values of γ below 0 are not prohibited in the model, but those individuals would prefer to invest a negative amount of capital into production, so they are not entrepreneurs.

$$P\left(z > \underbrace{\left(e^{\delta_S+\epsilon_S}\left(\frac{\alpha\gamma}{r}\right)^{\frac{-\alpha}{1-\alpha}}\theta^{\rho_S-\frac{1}{1-\alpha}}+\alpha\gamma-1\right)\sqrt{\theta^{\frac{1}{1-\alpha}}\left(\frac{\alpha\gamma}{r}\right)^{\frac{\alpha}{1-\alpha}}}}_{\text{RHS}}\right) \tag{1.8}$$

As the RHS changes, this affects the probability that entrepreneurship is the better option. When the RHS is 0, the odds are 50-50 that V_E turns out to be greater than V_S. Increasing the RHS decreases the probability of entrepreneurship, and decreasing it does the opposite. We can visualize changes in the RHS as movements around the support underneath a surface described by the CDF and observing how the height of surface changes. Of particular interest are the gradients of this surface with respect to θ and γ.

So far we have not considered whether any of the elements that compose the RHS of Equation 1.8 are at all related to each other. Implicitly, we have assumed that they are not. Relations between variables are important because they affect the value of the total derivative with respect to any variable that is not independent of some other variable. If all variables are independent, then all total derivatives equal all partial derivatives. That is, when I go along the surface in the direction of increasing variable x, the value of variable y remains constant; it does not change as x changes. In other words, the impact on the level of the surface occurs only through the direct impact of changing x. If instead y is linked in some way to x, then it cannot remain constant as x changes. The impact on the level then occurs both through the direct channel of x changing and the indirect channel of y changing simultaneously.

Remember that the values we observe in Figure 1.1 show a total derivative. To reconcile the puzzling observation with the predictions of the LR model, we need to extend the model to allow for a relationship between cognitive skill and risk tolerance.

1.4.3 Extending the model

I first establish the predictions of the model in the case when θ and γ are independent. In all cases I consider all other parameters to be determined independently of these two characteristics.

To see how the value of the RHS in Equation 1.8 changes with θ, I plot[6] the values of the RHS as a function of θ. Specifically, I do this relative to the indifference cut-off, subtracting its value from the RHS, as the z cut-off also changes with θ. This gives the

[6]The Desmos online graphing calculator is a useful tool for this.

relevant measure of where in the distribution one is, how far away one is from indifference. For the remainder of the section, RHS will be used to mean this "net" RHS.

When θ is independent of all other parameters, the RHS steadily decreases in θ. This means that the probability of entrepreneurship is predicted to steadily increase, tracing out the full standard normal CDF as θ increases to infinity. As θ is in practice limited, we would expect to observe just the initial part of this CDF. This is the benchmark shape for Figure 1.1 that was discussed in Section 1.3.

The shape of the relationship between the RHS and θ is not qualitatively affected by changes in the other parameters. As a reasonable starting point, I set $\delta_S + \epsilon_S = 0$, $\alpha = .33$, $\rho_S = .75$, and $r = 1.2$. Varying the value of γ from 0 to 1 or changing the other parameters still results in the RHS decreasing with θ, as we would expect.

Performing the same exercise with γ yields the same qualitative impact on the RHS – a steady decrease. Both of these results make sense, as we expect the direct impact of each characteristic to increase the probability of entrepreneurship in the model. The decreasing slope with respect to θ is less steep than that with respect to γ. However, the support of θ in the model is $[0, \infty)$, so it is not clear if here we can interpret θ as having a lower direct impact on the probability of entrepreneurship. For this we would need a maximum value of θ. Such a value certainly exists, although it is not specified in the LR model. The empirical evidence I present in Section 1.5 is consistent with ability having a lower direct impact on the likelihood of pursuing entrepreneurship than does risk tolerance.

I extend the LR model by linking the values of θ and γ, forcing them to be jointly determined, and then observe how the value of the RHS behaves when either value is changed. I use the simplest functional form that accounts for the key nonlinearity described in Figure 1.4[7]. Risk tolerance and ability increase together at low values of risk tolerance, and then ability decreases as risk tolerance continue to increase.

$$\gamma(\theta) = \begin{cases} n(\theta - m) \\ 1 - n(\theta - m) \end{cases} \qquad \theta(\gamma) = \begin{cases} \frac{1}{n}\gamma + m & \text{when } \gamma \in (0, 0.5) \\ -\frac{1}{n}\gamma + (m + \frac{1}{n}) & \text{when } \gamma \in [0.5, 1] \end{cases}$$

Linking θ and γ imposes a restriction on the values θ can take. With the above functional form, the lowest value of θ is m, which occurs when γ is either 0 or 1, and the highest

[7] One can add an error term to allow for a noisy nonlinear relationship between the two quantities, but for the following analysis it is an unnecessary complication.

value is $\theta = m + \frac{1}{2n}$, which occurs when $\gamma = 0.5$. I set $m = 2$ and $n = 0.3$, but again, changing these values does not alter the qualitative behavior of the RHS described below.

First consider how the RHS varies with γ, the level of risk tolerance, when we allow a changing γ to affect cognitive skills. As γ increases from 0 to 0.5, the RHS decreases, so the probability of entrepreneurship increases in risk tolerance. As γ continues to increase, the same generally continues to hold in most cases. It is possible to observe a U-shaped RHS between 0.5 and 1 if ρ_S is very close to 0 (ability is not useful in salaried employment), if α is very high (e.g. 0.95), or if $\delta_S + \epsilon_S$ is very high (e.g. 2.5, indicating that nonpecuniary benefits of salaried employment relative to entrepreneurship are high and/or the individual has very high salaried-employment-specific skills). Generally, however, the RHS steadily decreases as risk tolerance increases, meaning the probability of entrepreneurship increases as well. This is consistent with what we observe in figure 1.2.

To understand the impact of θ on the probability of entrepreneurship, we need to consider the behavior of the RHS in two cases, when γ is low and when it is high. In the first case, as θ increases, this is associated with an increase in γ as well. Both direct effects reinforce each other. The RHS decreases, so the model predicts that the probability of entrepreneurship will increase with cognitive skill.

In the second case, the opposite occurs. When γ is high, increasing θ decreases γ, overwhelming the impact of higher ability. The RHS increases, so now the model predicts that the probability of entrepreneurship will decrease with cognitive skill.

The overall marginal effect of increasing θ on the probability of entrepreneurship will be the simple average of the two cases above. This is because by construction, each level of θ corresponds to one low value of γ and to one high value of γ. Figure 1.7 shows the change in the RHS that results from increasing θ:

At low levels of θ, increasing it drops the value of the RHS – the probability of entrepreneurship increases. However, this soon levels off, and the value of the RHS remains flat as θ continues to increase. The probability of entrepreneurship thus remains approximately constant beyond middle levels of ability. Once again, this result is robust to changes in other parameter values. The pattern in Figure 1.7 is exactly what we observe in Figure 1.1: those with the highest cognitive skills are about as likely to pursue entrepreneurship as are those with average skills.

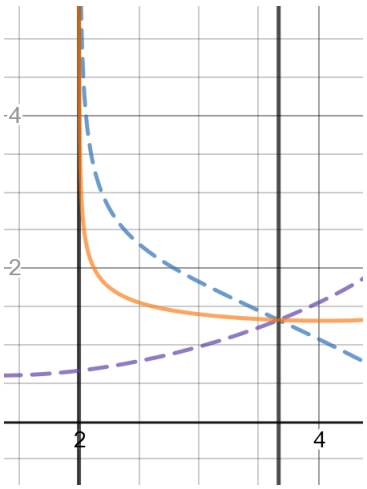


Figure 1.7: Impact of ability on the "net" RHS of Equation 1.8 when ability and risk tolerance are not independently determined. The two vertical black lines show the minimum and maximum values of θ. The decreasing dotted blue line shows the low-γ case. The increasing dotted purple line shows the high-γ case. The solid orange line is the average of the two.

1.4.4 Nudging into entrepreneurship

The model predicts that the direct impact of risk tolerance on the likelihood of pursuing entrepreneurship is positive. That is, if an individual's risk tolerance were to increase, she would be more likely to start her own incorporated business. The risk tolerance measure γ combines within it two concepts, the individual's attitude towards risk, τ, and the "riskiness" of entrepreneurship, σ^2. Thus far we have disregarded the latter, treating it as constant. However, while it may be difficult or impossible for a policy-maker to influence an individual's τ, it is much more feasible to shift γ upwards by lowering σ^2, in other words, providing support for a budding business.

Figure 1.5 shows us that most entrepreneurship occurs when individuals rate their level of risk tolerance as 5 or higher – the high-γ portion of the distribution, where ability and risk tolerance are inversely related. If we were to uniformly increase γ by making entrepreneurship safer, the individuals who now choose to become entrepreneurs should on average be of higher ability than those who already chose to incorporate with their original value of γ. Loosening the γ constraint should thus allow people with a higher τ to enter, and these people should on average have a higher θ as well.

An optimal way to determine whether it is possible to nudge high ability individuals into incorporated entrepreneurship would be through a randomized controlled trial. The NLSY data do not provide enough power for a detailed analysis using a natural experiment, but I present some basic observations.

Effect of riskiness on entry into entrepreneurship

NLSY79	logit FE		OLS FE	
Exit rate	-.0749551***	-.2460391***	-.0006184***	-.0019191***
	(.0174027)	(.0259045)	(.0001377)	(.0002006)
Controls	No	Yes	No	Yes
Number individuals	796	760	796	760
Number observations	14,658	13,664	14,658	13,664
NLSY97	logit FE		OLS FE	
Exit rate	-.0616624**	-.1189491**	-.0007525**	-.0007357
	(.0271838)	(.0479704)	(.0003017)	(.000497)
Controls	No	Yes	No	Yes
Number individuals	559	450	559	450
Number observations	5,973	3,878	5,973	3,878

Table 1.4: A higher establishment exit rate in a given industry and year is associated with a lower probability of incorporated entrepreneurship. Regressions reported above include an individual fixed effect. Also reported are the number of marginal individuals who switch in/out of entrepreneurship at some point during the study, on whom the estimates are based. Controls included are income in the previous year (in thousands of 2010 dollars) and industry dummies, as individuals can change industries. Income has a very small but statistically significant positive effect on entrepreneurship participation. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

As a measure of "riskiness" of starting a business, I use data on the rate of establishments in a particular industry exiting the market in a given year from the United States Census Bureau. This value is based on data gathered across the country, so a single individual would not be able to influence the measured rate. The rate of exit a person observes in their industry each year is thus taken as given. If this measure is an adequate proxy for the business climate, it should be associated with changes in the rate of entrepreneurship in the NLSY samples. This is indeed the case. In a fixed effects logistic regression of the probability of being an entrepreneur on the rate of exit, an increase in the exit rate is associated with a lower likelihood of entrepreneurship. The fixed effects control for individual-level characteristics that do not vary over time, including ability and risk tolerance, and the estimates are based off of those marginal individuals who switch into or out of entrepreneurship during the course of the study. I also report the results of OLS estimation with individual fixed effects, and results including controls for the individual's previous year's earnings and their industry, as people do change industries over the course of their careers. It remains the case that the establishment exit rate is inversely associated with entry into entrepreneurship.

As people enter and leave entrepreneurship, the characteristics of the pool of entrepreneurs vary each year. We can see how changes in the "riskiness" of the business climate (i.e. changes in σ^2) affect the average AFQT and the average risk tolerance of those who choose

to be entrepreneurs that year.

While these results are only indicative, it is interesting to see that as riskiness of running a business increases, so does the average risk tolerance of entrepreneurs (though not significantly). Meanwhile the opposite holds true for ability. Thus, it is possible that by decreasing the risk associated with running one's own business, a greater number of highly able people would enter, but further research is necessary.

Changes in pool of entrepreneurs with increasing riskiness

	AFQT	Risk tolerance
NLSY79	-2.0198***	.0203394
	(.3698833)	(.0418038)
n	1,528	1,313
NLSY97	-1.009317*	.0142251
	(.6054531)	(.0450683)
n	824	1,005

Table 1.5: Regression of AFQT and of risk tolerance on the establishment exit rate, only across those person-year observations in which the person is incorporated. Increasing exit rate (riskiness) seems to decrease the average AFQT of entrepreneurs in both samples, while increasing (not significantly) the average risk tolerance. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

1.5 Model estimation and counterfactuals

1.5.1 Specification

To estimate the model, I start with the indifference expression in Equation 1.5. As discussed in section 1.4.2, this defines the shape of the contour lines of the $P(V_E > V_S)$ surface in the θ-γ space. Along the curve described by the equation, the probability that entrepreneurship produces greater utility than salaried employment remains a constant value. If we change θ or γ and move off of this curve, the probability changes to a different value – but now we find ourselves on a different indifference curve, which will still have the same kind of shape.

By taking the natural log of both sides, we see that the expression becomes linear in $\ln\theta$ and $\ln\gamma$:

$$\left(\rho_S - \frac{1}{1-\alpha}\right)\ln\theta = \left[\frac{\alpha}{1-\alpha}\ln\left(\frac{\alpha}{r}\right) + \ln(1-\alpha) - \delta_S - \epsilon_S\right] + \left(\frac{1}{1-\alpha}\right)\ln\gamma \qquad (1.9)$$

This means that in the transformed space, the indifference curves are linear, and we can use a simple probit to estimate the surface above the $\ln\theta$-$\ln\gamma$ plane. Probit is appropriate because the shape the model predicts is a normal CDF, as ν is distributed normally.

So far we have established that the probability of entrepreneurship is determined by a linear combination of $\ln\theta$ and $\ln\gamma$:

$$P(Y_{it} = 1|\theta_i, \gamma_i) = \Phi(\beta_0 + \beta_1 \ln\theta_i + \beta_2 \ln\gamma_i)$$

Estimating this directly, however, does not take into account the nonlinear relationship between ability and risk tolerance. In order to be able to interpret the coefficients estimated by a linear regression model as partial derivatives (direct effects), we need to allow the correlation between the two characteristics to switch from positive to negative for low vs. high values of risk tolerance. If we were to drop risk tolerance from the model, this would impact the coefficient on ability in opposite ways depending on whether it is positively or negatively correlated with risk tolerance – and this depends on the level of risk tolerance. A simple change is needed:

$$P(Y_{it} = 1|\theta_i, \gamma_i) = \Phi\{\beta_0 + \beta_1(\ln\theta_i \times \mathbb{1}[\gamma_i < \bar{\gamma}]) + \beta_2(\ln\theta_i \times \mathbb{1}[\gamma_i \geq \bar{\gamma}]) + \beta_3 \ln\gamma_i\} \quad (1.10)$$

Based on the observation in Figure 1.3, I choose 5 as the $\bar{\gamma}$ to use in estimation. To estimate this model, we need to take logs of ability and risk tolerance, so zero values are problematic. Fortunately, there is nothing fundamental about a value of 0 in either variable – it is an arbitrarily-chosen lowest point. I thus adjust the variables so the new lowest value is 1 before proceeding with the estimation.

1.5.2 Estimation

Table 1.6 reports the results. As in the discussion of the model, I ignore the unincorporated self-employed, combining them with the salaried. The table includes estimates for each cohort estimated in two different ways. The first keeps each observation at the person-year level, so the dependent variable is entrepreneurship status in each year. To control for changes to γ through σ^2, I include the establishment exit rate described earlier. I also include the individual's entrepreneurship status in the previous year, as entrepreneurship status is unlikely to be an i.i.d. process but instead depends on past experience. For these regressions the standard errors are clustered at the individual level.

Probit regression - selection into entrepreneurship

	NLSY79		NLSY97	
	Person-year	Person	Person-year	Person
AFQT	.0060408	-.0156233	-.0364986	.0281823
(low risk tolerance)	(.0256918)	(.0336081)	(.0423733)	(.0445010)
AFQT	.0353375	.0563882*	-.0080455	.0755322**
(high risk tolerance)	(.0245766)	(.0317610)	(.0374996)	(.0384426)
Risk tolerance	.0627717	.0990755*	.2971609**	.2651479**
	(.0490613)	(.0575332)	(.1230023)	(.1188887)
Entrepreneur previous year	1.990449***	-	2.225054***	-
	(.0613140)		(.0823398)	
Income in past year	.0017199***	-	.0050779***	-
(2010 $1000s)	(.0002807)		(.0008786)	
Estab. exit rate	-.0522084***	-	-.0109014	-
("riskiness")	(.0123320)		(.0223669)	
Childhood family income	.0017718***	.0020117***	.0004714	.0007240
(2010 $1000s)	(.0004696)	(.0006562)	(.0004785)	(.0005521)
Two-parent HH	-.0528674	-.0999683*	-.0545169	-.0098277
	(.0381971)	(.0525844)	(.0547383)	(.0611085)
Mother's education	.0261786***	.0428824***	.0031917	.0029113
	(.0089845)	(.0115869)	(.0115481)	(.0126662)
Father's education	-.0040971	.0014876	.0227115**	.0123596
	(.0066316)	(.0087721)	(.0112107)	(.0123914)
Black	-.1594923***	-.2433696***	.0882053	.0680362
	(.0498651)	(.0687275)	(.0740262)	(.0807415)
Hispanic	-.0436163	.0189614	.1135161	.1262783
	(.0547214)	(.0736877)	(.0739255)	(.0815494)
Female	-.0975346**	-.2570386***	-.1467682***	-.2547798***
	(.0398873)	(.0481688)	(.0545605)	(.0579082)
Constant	-2.07053***	-1.998738***	-2.918532***	-2.312771***
	(.1995637)	(.1689988)	(.4212323)	(.3066065)
n	103,287	5,779	30,524	4,264

Table 1.6: Marginal effects on entry into incorporated self-employment. Standard errors clustered at the individual level (columns 1 and 3) or robust (columns 2 and 4). In columns 1 and 3, the dependent variable is entrepreneurship status in a given year, and in columns 2 and 4 it is having ever been an entrepreneur across all years for which data is available. "AFQT" denotes the natural log of AFQT scores, and "Risk tolerance" denotes the natural log of risk tolerance values. The lowest value of each variable is 0. The cut-off for low vs high risk tolerance is 5 based on the original measurement. Columns 1 and 3 additionally control for the person's industry in a given year, although those results are not reported. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

The second set of estimates is closer to what we have modeled. The model does not say when a person would become an entrepreneur or for how long they would run their business, only that some individuals are more likely to pursue this career at some point than are others. I thus label as an entrepreneur each person who is an incorporated business owner at any point during the measured time period and estimate the model at the individual level. For these regressions standard errors are estimated robustly, allowing for potential heteroskedasticity or autocorrelation.

The following control variables are also included in the model: income earned the previous year in thousands of 2010 dollars, the family income when the respondent was a child in thousands of 2010 dollars, whether the respondent grew up in a two-parent household, the mother's and father's education, dummies for race and sex, and the industry in which the individual works. The variables that change over time are dropped from the person-level regressions.

We see that on average, ability has a positive direct impact on the probability of entrepreneurship, and that its impact tends to be higher at higher levels of risk tolerance. In the NLSY97 sample using person-year observations, the point estimate on ability is negative at both high and low levels of risk tolerance, although the coefficients are not significant. The marginal effect of risk tolerance is larger and positive in every specification, and is statistically significant in three of the four regressions. This is consistent with risk tolerance having a greater impact on whether an individual becomes an entrepreneur than does the person's cognitive ability.

In both samples, income earned the previous year has a statistically significant positive effect on entry. It is those who earn more who tend to become incorporated business owners. Prior entrepreneurship status predicts current entrepreneurship status well. This is not surprising, considering that most individuals are never entrepreneurs.

Being female makes one much less likely to be an incorporated entrepreneur in either cohort. Women do tend to be more risk averse in both samples, but this is an additional impact after controlling for risk preferences. This female gap in entrepreneurship participation is large, and is highly statistically significant in all specifications. At the average value of all other regressors, the probability of entrepreneurship is lower for a woman by 3.5 to 3.8 percentage points as compared to a man. This is huge, considering that the maximum predicted probability of entrepreneurship for any individual in either sample is 35%.

In one regression, having a more highly educated father is associated with higher probability of entrepreneurship, but this effect is not replicated in the other specifications. Growing up in a two-parent household also produces a negative, marginally significant coefficient in one specification, but in this case all regressions document a small negative effect.

Interestingly, while a higher mother's level of education is associated with greater pursuit of entrepreneurship in the NLSY79, this is not the case for the younger cohort. It is possible that educated women became more prevalent by the time the 1997 cohort was born, but that it made the family unusual in some way for the 1979 cohort.

Another potentially encouraging observation is that while a higher family income during the respondent's childhood is positively associated with entry into entrepreneurship in the NLSY79, it is no longer significant in the NLSY97. In addition, while being black is negatively and significantly associated with entrepreneurship in the NLSY79, the coefficient estimates are positive, although not significant, in the NLSY97.

1.5.3 Counterfactuals

For the counterfactual analysis, I focus on the person-level models, the estimates for which are reported in columns 2 and 4 of Table 1.6. The reported regression estimates can be thought of as describing a surface above a multi-dimensional support. This is the predicted probability of entrepreneurship at each combination of independent variable values. I am interested to see how this surface varies across the support defined by only two of the variables, risk tolerance and ability. This means I need to set the values of all other variables to some constants. The natural choice is the average value of each variable in the sample used in estimation. The resulting object is then a normal CDF-shaped surface above the $\ln\theta$-$\ln\gamma$ plane, gradually increasing as $\beta_1 \ln\theta + \beta_2 \ln\gamma$ increases[8]. Choosing a different set of constants for the control variables is analogous to shifting this CDF "backwards" or "forwards" relative to its contour lines, bringing high probability values closer to or further from the origin.

In order to perform calculations based on the estimated surface, I discretize the $\ln\theta$-$\ln\gamma$ space, separating it into bins as in Figure 1.3. I then take the predicted probability of entrepreneurship at the middle values of each bin and apply this value for the whole bin.

[8]Technically, the estimated specification defines two different CDFs, one of which is defined on low levels of γ and the other of which is defined on high levels of γ.

The surface described above is a point estimate, a $\Phi(\hat{y})$, where $\hat{y}$ depends on the estimated values of the β parameters. I calculate the variance of $\hat{y}$ in each $(\ln\theta, \ln\gamma)$ bin by pre- and post-multiplying the estimated variance-covariance matrix of the $\hat{\beta}$ by a vector of regressor values. I keep the value of the control variables set to their in-sample average value and I change the values of $\ln\theta$ and $\ln\gamma$ as appropriate for each bin. I then take the square root of the variance to get a standard error and form a confidence interval around each $\hat{y}$ by adding/subtracting 1.96 times the standard error. Transforming the resulting $\hat{y}$s using the normal CDF then produces error bands for the surface.

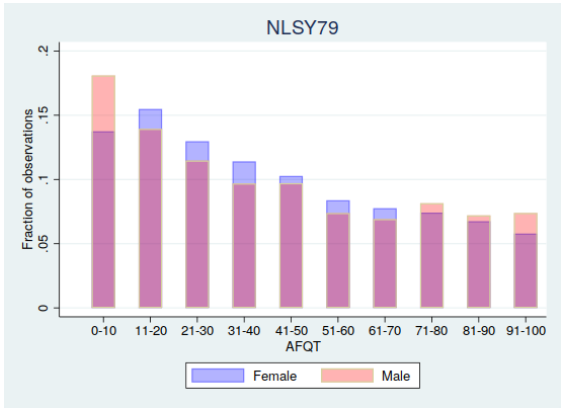
For each combination of risk tolerance and ability, we now have a prediction for how likely an individual with those characteristics is to become an entrepreneur. Within each bin, we expect to observe $n \cdot p$ entrepreneurs, where n is the number of people whose values of ability and risk tolerance fall into that bin, and p is the predicted probability of entrepreneurship for the bin. Summing the expected number of entrepreneurs across all bins in the sample gives us the total predicted number of entrepreneurs. We can get error bands on this number by performing the calculation using the error bands on the probability of entrepreneurship.

The actual number of individuals who are at any point observed to be entrepreneurs is 843 in the NLSY79 and 663 in the NLSY97. The model predicts 632.19 entrepreneurs in the NLSY79 with a confidence interval of (-99.78, 1364.16). In the NLSY97, the total predicted number is 485.01 with a confidence interval of (-273.22, 1243.24). These values underpredict, but this is likely because the values chosen for the control variables bias the prediction downward deterministically. In all cases considered, the model predicts approximately 75% of the actual number of entrepreneurs. We can thus still use the predictions to gain insight about relative changes under various counterfactual scenarios.

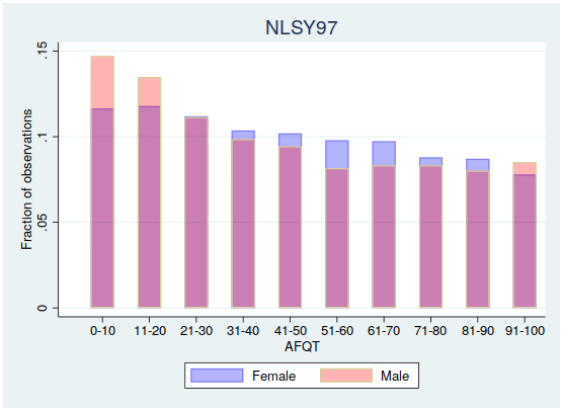
Gender gap in entrepreneurship

There is a notable gap in the number of male and female entrepreneurs in the data. In the NLSY79, the gender breakdown of entrepreneurs is 310 women and 533 men. In the NLSY97, the numbers are 237 and 426. In both cases, approximately 36% of entrepreneurs are women. There is evidence in the literature that women are more risk averse (for example, Borghans et al., 2009, Agnew et al., 2008), so that could be the reason for the difference. Figure 1.8 shows that women and men do differ in their risk tolerance as well as cognitive ability in the NLSY cohorts. Men are relatively more likely to obtain a very low or very high score on the AFQT, and men are also relatively more likely to be

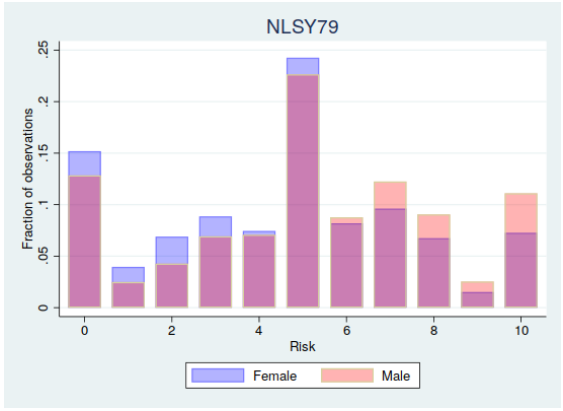
highly risk tolerant. Perhaps it is this relative over-representation in the right tail of both distributions that leads men to become entrepreneurs more frequently. Maybe women are just less willing to take the necessary risks.



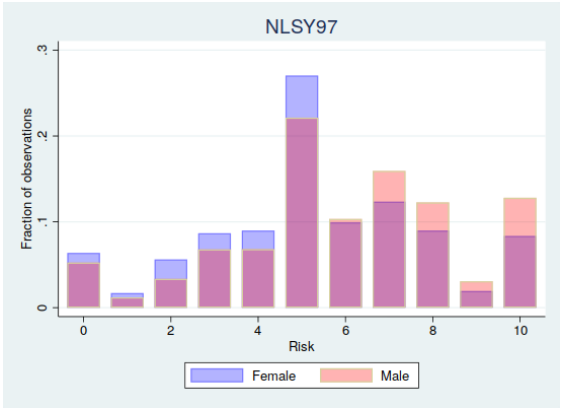
(a) AFQT, NLSY79



(b) AFQT, NLSY97



(c) Risk tolerance, NLSY79



(d) Risk tolerance, NLSY97

Figure 1.8: Distribution of ability and risk tolerance by sex

That explanation seems improbable. In both model specifications estimated, the coefficient on the female dummy is negative and highly statistically significant. This shows that even controlling for differences in ability and risk tolerance, women are still at a disadvantage when it comes to pursuing entrepreneurship. The source of this "handicap" is important to investigate in future research. It could be due to cultural norms or greater difficulty obtaining funding or perhaps other factors. Here, I calculate the extent to which it inhibits entrepreneurship among women, and to what extent it is women's risk tolerance or ability that prevent them from becoming entrepreneurs.

So far, the surface we have used to calculate the number of entrepreneurs is based on setting the value of the female dummy to the average in the sample. There are approximately the same number of men and women in either NLSY cohort, so this value is roughly 0.5. As discussed at the beginning of this section, changing the value of the control variables shifts the probability distribution. For men, changing the dummy to 0 pulls the surface

closer to the origin relative to the surface used thus far. For women, setting it to 1 pushes the surface further away from the origin. This has the effect of raising the probability of entrepreneurship in every bin for men, and lowering it for women. We can thus consider two scenarios. How many more female entrepreneurs would we expect to observe if there were no "woman handicap" and both men's and women's probabilities of entrepreneurship were determined by the men's surface? Alternatively, if we keep the surface fixed for both genders, to what extent does the model predict a gender gap in entrepreneurship, and how does it change if women's characteristics are reweighed to match those of men?

First I look at the number of male and female entrepreneurs the model predicts if we set the female dummy to 0 and 1, respectively, shifting the surface accordingly for each gender. In the NLSY79, the model predicts 249.71 (CI -179.04, 678.46) female entrepreneurs and 399.40 (CI 11.00, 787.80) male entrepreneurs. In the NLSY97, it is 182.27 (CI -257.00, 621.53) female entrepreneurs and 314.19 (CI -90.61, 718.99) male entrepreneurs. The fraction of predicted entrepreneurs who are women closely matches the true value, 38% in the NLSY79 and 37% in the NLSY97.

Reweighing the women's joint distribution of risk tolerance and ability to match that of men while keeping the "handicap" in place brings the number of female entrepreneurs up, but only slightly. The new predicted numbers are 266.21 (CI -164.09, 696.51) in the NLSY79 and 195.97 (CI -237.16, 629.11) in the NLSY97. Accounting for women's differing ability and risk preferences relative to men thus brings the fraction of female entrepreneurs up to 40% and 38%, respectively.

The change is more drastic if we remove the gender bias while keeping women's risk preferences and ability as they actually are distributed. In this case, both men's and women's probability of entrepreneurship is determined by the men's surface. Now, the model predicts 398.67 (CI -12.96, 810.31) female entrepreneurs in the NLSY79 and 292.35 (CI -117.79, 702.48) female entrepreneurs in the NLSY97. If we additionally reweigh women's risk preferences and ability to match those of men, we get 422.33 (CI 11.63, 833.02) and 312.17 (CI -90.03, 714.36) female entrepreneurs in each cohort.

This last case amounts to parity, the number of female entrepreneurs we would expect to see if risk preferences and ability were the same for men and women, and if there were no other bias preventing women from becoming entrepreneurs. The fraction of women who become entrepreneurs is then equal to the fraction of men who become entrepreneurs. Allowing for gender differences in ability and risk tolerance decreases the number of female entrepreneurs by 5.6% and 6.3% in the NLSY79 and NLSY97, respectively. Allowing for

the "woman handicap" decreases it by a further 35.3% in both cohorts. Differences in risk preferences or ability thus explain approximately 13.7%-15.3% of the gender gap in entrepreneurship.

Lowering risk for high ability

As noted previously, high cognitive ability is associated with middling levels of risk tolerance. At the same time, the marginal effect of greater risk tolerance on entry into entrepreneurship is relatively high, and the marginal effect of ability is higher when risk tolerance is high. It is thus interesting to consider the impact a risk-lowering intervention may have on those with high cognitive skill. This is essentially the service that some business incubators offer to potential start-up founders.

The number of entrepreneurs in the NLSY79 and NLSY97 whose AFQT score is 70 or higher is 215 and 151, respectively. Using the model with average values of all control variables, the predicted number of such high-ability entrepreneurs is 132.07 (CI -18.87, 283.00) in the NLSY79 and 131.62 (CI -51.59, 314.84) in the NLSY97.

I consider an intervention which lowers the risk involved in running an incorporated business by an amount equivalent to shifting individuals up one level of risk tolerance, keeping all else constant. The new predicted number of highly able entrepreneurs is 142.85 (CI -2.41, 288.12) in the NLSY79 and 145.82 (CI -29.74, 321.38) in the NLSY97. The first is equivalent to an 8.16% increase, and the second is a 10.79% increase. In practice, we would need to determine the appropriate measures which lower the risks involved, but lowering this barrier to entry could potentially encourage substantial entry into entrepreneurship among high-ability individuals.

1.6 Conclusion

Governments around the world are keen to encourage entrepreneurship, seeing it as a key source of innovation and of economic growth. As always, it comes down to individual choice – who will respond to government incentives? What kind of businesses would they start? Are different incentives necessary to encourage the would-be founders of an ambitious, transformative start-up as compared to a local small family business? Based on past research and the NLSY data, it appears personal characteristics of potential entrepreneurs play a nontrivial role in determining who decides to start a business and who prefers

the comfort of a stable job with an employer. In particular, if it is the formation of innovative technological companies that a government would like to stimulate, the pool of potential founders must be composed of individuals with a great deal of technical knowledge and intellectual ability. At the same time, entrepreneurship of any kind is an uncertain endeavour that requires a sufficient tolerance for risk. The combination of both characteristics in one person is rare. Through taking steps to lower the risk associated with founding one's own company, it may be possible to nudge a greater number of high ability individuals into entrepreneurship, potentially leading to greater innovation and commercialization of the new technologies. There is evidence that it is possible to induce people to engage in innovation (Zivin and Lyons, 2018), so it is likely incorporated entrepreneurship can also be induced by targeting those frictions which inhibit it.

Another longer-term strategy to encourage greater innovation and entrepreneurship is to invest in people – after all, individuals' personal traits may either help or hinder their probability of becoming an entrepreneur. While it is relatively rare to find an adult who is both highly intelligent as well as sufficiently risk tolerant, this does not have to be true for the next generation. Children are in the process of developing both their cognitive and noncognitive abilities throughout their childhood. As has been previously shown by Heckman and Masterov (2007), if governments insufficiently invest in the development of young children growing up in disadvantaged environments, they miss out on an investment with high economic and social returns. Perhaps one of these returns is a greater number of entrepreneurs. Bell et al. (2019) have found that exposure to innovation increases children's propensity to innovate themselves as adults. Being allowed to develop a healthy attitude towards risk as well as their cognitive abilities may do the same for children's future entrepreneurial careers. In both waves of the NLSY, a lower family income in childhood is associated with lower performance on the AFQT. Its impact on risk tolerance is dual – those who grow up with fewer resources appear to either become very risk averse or very risk tolerant. Investing in underprivileged children may thus change the joint distribution of the two characteristics compared to what is currently observed in adults, and lead to both greater innovation and entrepreneurship in society.

Exploring the potential for inducing a greater number of individuals to become entrepreneurs is left for future research, together with other important considerations, such as how the performance of these marginal individuals would differ from that of those who would enter anyway, and whether the welfare trade-offs are worthwhile. A randomized controlled trial could go a long way towards answering these policy-relevant questions. If highly intelligent people are capable of developing a new technology and bringing it to market, but the

perceived riskiness of attempting to do this prevents them from acting, then lowering that barrier can be an effective way to encourage the formation of such companies. However, it is also possible that those who do not start a business choose to do so because they know they cannot build a company, in which case convincing them otherwise will not be effective at producing greater innovation. This is a difficult assessment for an individual to make – after all, many factors can influence whether a company succeeds or fails. It is much easier to determine whether a person has the underlying technical knowledge necessary to try. Through differentiated recruitment messaging that emphasizes low or high levels of risk involved (e.g. "80% of our founders found the experience worthwhile" vs. "1% of our founders have sold their company for more than £1 million"), a business incubator could determine whether marginal individuals are self-selecting efficiently. If they are not, lowering the relevant barrier may result in greater innovation, growth, and potentially higher welfare.

References

Julie R. Agnew, Lisa R. Anderson, Jeffrey R. Gerlach, and Lisa R. Szykman. Who chooses annuities? An experimental investigation of the role of gender, framing, and defaults. *The American Economic Review*, 98(2):418–422, 2008.

Taehyun Ahn. Attitudes toward risk and self-employment of young workers. *Labour Economics*, 17(2):434–442, 2010.

Thomas Åstebro, Holger Herz, Ramana Nanda, and Roberto Weber. Seeking the roots of entrepreneurship: Insights from behavioral economics. *Journal of Economic Perspectives*, 28(3):49–70, 2014.

William Baumol. Entrepreneurship: Productive, unproductive, and destructive. *The Journal of Political Economy*, 98(5):893–921, 1990.

Alex Bell, Raj Chetty, Xavier Jaravel, Neviana Petkova, and John Van Reenen. Who becomes an inventor in America? The importance of exposure to innovation. *The Quarterly Journal of Economics*, 134(2):647–713, 2019.

Daniel Benjamin, Sebastian Brown, and Jesse Shapiro. Who is 'behavioral'? Cognitive ability and anomalous preferences. *Journal of the European Economic Association*, 11 (6):1231–1255, 2013.

Lex Borghans, James J. Heckman, Bart H. H. Golsteyn, and Huub Meijers. Gender differences in risk aversion and ambiguity aversion. *Journal of the European Economic Association*, 7(2-3):649–658, 2009.

Matthias Brachert and Walter Hyll. On the stability of preferences: Repercussions of entrepreneurship on risk attitudes. *SOEP Papers no. 667*, 2014.

Marco Caliendo, Frank Fossen, and Alexander Kritikos. Risk attitudes of nascent entrepreneurs – new evidence from an experimentally validated survey. *Small Business Economics*, 32(2):153–167, 2009.

David de Meza and Clive Southey. The borrower's curse: Optimism, finance and entrepreneurship. *The Economic Journal*, 106(435):375–386, 1996.

Thomas Dohmen, Armin Falk, David Huffman, Uwe Sunde, Jürgen Schupp, and Gert Wagner. Individual risk attitudes: New evidence from a large, representative, experimentally-validated survey. *IZA discussion paper no. 1730*, 2005.

Thomas Dohmen, Armin Falk, David Huffman, and Uwe Sunde. Are risk aversion and impatience related to cognitive ability? *American Economic Review*, 100(3):1238–1260, 2010.

Thomas Dohmen, Armin Falk, David Huffman, Uwe Sunde, Jürgen Schupp, and Gert Wagner. Individual risk attitudes: Measurement, determinants, and behavioral consequences. *Journal of the European Economic Association*, 9(3):522–550, 2011.

Thomas Dohmen, Armin Falk, Bart Glosteyn, David Huffman, and Uwe Sunde. Risk attitudes across the life course. *The Economic Journal*, 127(605):F95–F116, 2017.

David Evans and Boyan Jovanovic. An estimated model of entrepreneurial choice under liquidity constraints. *Journal of Political Economy*, 97:808–827, 1989.

David Evans and Linda Leighton. Some empirical aspects of entrepreneurship. *American Economic Review*, 79(3):519–535, 1989.

Robert Fairlie. Drug dealing and legitimate self-employment. *Journal of Labor Economics*, 20(3), 2002.

Armin Falk, Anke Becker, Thomas Dohmen, Benjamin Enke, David B. Huffman, and Uwe Sunde. The nature and predictive power of preferences: Global evidence. *IZA discussion paper no. 9504*, 2015. (Quarterly Journal of Economics 2018).

Shane Frederick. Cognitive reflection and decision making. *Journal of Economic Perspectives*, 19(4):25–42, 2005.

Matteo Galizzi, Sara Machado, and Raffaele Miniaci. Temporal stability, cross-validity, and external validity of risk preferences measures: experimental evidence from a UK representative sample. *The London School of Economics and Political Science, Department of Social Policy, London, UK*, 2016. URL `http://eprints.lse.ac.uk/67554/`.

Carole Gallagher. Reconciling a tradition of testing with a new learning paradigm. *Educational Psychology Review*, 15(1):83–99, 2003.

Nicola Gennaioli, Rafael La Porta, Florencio Lopez de Silanes, and Andrei Shleifer. Human capital and regional development. *Quarterly Journal of Economics*, 128:105–164, 2013.

Barton Hamilton. Does entrepreneurship pay? An empirical analysis of the returns to self-employment. *Journal of Political Economy*, 108:604–631, 2000.

James J. Heckman and Dimitriy V. Masterov. The productivity argument for investing in young children. *Review of Agricultural Economics*, 29(3):446–493, 2007.

Rebecca Henderson and Kim Clark. Architectural innovation: The reconfiguration of existing product technologies and the failure of established firms. *Administrative Science Quarterly*, 35(1):9–30, 1990.

Hakan Holm, Sonja Opper, and Victor Nee. Entrepreneurs under uncertainty: An economic experiment in China. *Management Science*, 59(7), 2013.

Erik Hurst and Benjamin Wild Pugsley. What do small businesses do? *Brookings Papers on Economic Activity*, 2:73–118, 2011.

Hans Hvide and Georgios Panos. Risk tolerance and entrepreneurship. *Journal of Financial Economics*, 111(1):200–223, 2014.

Anika Josef, David Richter, Gregory Samanez-Larkin, Gert Wagner, Ralph Hertwig, and Rui Mata. Stability and change in risk-taking propensity across the adult life span. *Journal of Personality and Social Psychology*, 2016.

Frank Knight. *Risk, Uncertainty, and Profit*. Houghton Mifflin, 1921.

Edward Lazear. Balanced skills and entrepreneurship. *American Economic Review*, 94 (2):208–211, 2004.

Irwin Levin, Stephanie Hart, Joshua Weller, and Lyndsay Harshman. Stability of choices in a risky decision-making task: A 3-year longitudinal study with children and adults. *Journal of Behavioral Decision Making*, 20(3):241–252, 2007.

Ross Levine and Yona Rubinstein. Smart and illicit: Who becomes an entrepreneur and do they earn more? *The Quarterly Journal of Economics*, 132(2):963–1018, 2017.

Ross Levine and Yona Rubinstein. Selection into entrepreneurship and self-employment. *NBER Working Paper No. 25350*, 2018.

Bruno Moreira, Raul Matsushita, and Sergio Da Silva. Risk seeking behavior of preschool children in a gambling task. *Journal of Economic Psychology*, 31(5):794–801, 2010.

Tobias Moskowitz and Annette Vissing-Jørgensen. The returns to entrepreneurial investment: A private equity premium puzzle? *American Economic Review*, 92:745–778, 2002.

Kevin Murphy, Andrei Shleifer, and Robert Vishny. The allocation of talent: Implications for growth. *Quarterly Journal of Economics*, 106(2):503–530, 1991.

William Nordhaus. Schumpeterian profits in the American economy: Theory and measurement. *NBER Working Paper No. 10433*, 2004.

Jörg Oechssler, Andreas Roider, and Patrick Schmitz. Cognitive abilities and behavioral biases. *Journal of Economic Behavior & Organization*, 72:147–152, 2009.

David Paulsen, Michael Platt, Scott Huettel, and Elizabeth Brannon. Decision-making under risk in children, adolescents, and young adults. *Frontiers in Psychology*, 2:72, 2011. URL `https://doi.org/10.3389/fpsyg.2011.00072`.

Claudia Sahm. How much does risk tolerance change? *Quarterly Journal of Finance*, 2 (4), 2012.

Henry Sauermann. Fire in the belly? Employee motives and innovative performance in start-ups versus established firms. *Strategic Entrepreneurship Journal*, 2017.

Hannah Schildberg-Hörisch. Are risk preferences stable? *Journal of Economic Perspectives*, 32(2):135–154, 2018.

Joseph Schumpeter. The creative response in economic history. *The Journal of Economic History*, 7(2):149–159, 1947.

Stephanie Schurer. Lifecycle patterns in the socioeconomic gradient of risk preferences. *Journal of Economic Behavior & Organization*, 119:482–495, 2015.

Jerome Vandenbussche, Philippe Aghion, and Costas Meghir. Growth, distance to frontier and composition of human capital. *Journal of Economic Growth*, 11:97–127, 2006.

Joshua Zivin and Elizabeth Lyons. Can innovators be created? experimental evidence from an innovation contest. *NBER Working Paper No. 24339*, 2018.

Chapter 2

Nonlinearities in 2SLS

2.1 Introduction

Mullainathan and Spiess (2017) point out that the first stage of the two stage least squares method is effectively a prediction task as opposed to an exercise in parameter estimation. The marginal effects we estimate are only a means to an end, a way to form a prediction $\hat{X}$, which we then use in the second stage to estimate the parameters that *are* of interest to us. We often ignore the complexities of the process generating the first stage relationship and model it linearly, relying on the many results that assure us this is a robust method. However, if the true relationship is not linear, Newey (1990) shows that it is possible to improve the efficiency of our parameter estimates by changing how we estimate the first stage – the authors emphasize this is where machine learning methods can be helpful.

In this paper, we evaluate the extent to which second stage estimates improve under several different nonlinear first stage data generating processes (DGPs). We find that the greatest gain in both efficiency and bias of the estimates occurs when the nonlinear DGP has an average slope close to 0 over the input domain. In this case, the instrument typically appears weak to the linear method, and as expected, it performs poorly, whereas a nonlinear approach is able to discern the true first stage relationship and produces accurate second stage parameter estimates. Alternatively, the instrument may appear strong when estimated on the drawn data sample – but it incorrectly identifies the underlying DGP and results in poor second stage estimates, potentially leading to mistaken inference.

The method we apply to estimate a nonlinear first stage is neural networks. This allows us to remain agnostic with regards to the shape of the relationship, but to ensure that we

allow enough flexibility to trace out any potential DGP. Unlike most papers which develop and analyze nonlinear instrumental variables methods (e.g. Newey et al., 1999; Horowitz, 2011), the second stage remains linear, as this is the most intuitive relationship commonly used in applied work. Without complicating the analysis of the question of interest, there is an advantage to more carefully considering the nature of the first stage.

It is important to emphasize that the set-up we propose is not, in fact, a "forbidden regression" (Hausman, 1975). That may have been so were we to use the predicted values as a regressor in the second stage (i.e. $\hat{\beta} = (\hat{X}'\hat{X})^{-1}\hat{X}'Y$). Instead, we use the predicted values as an instrument in the second stage, circumventing that problem (i.e. $\hat{\beta} = (\hat{X}'X)^{-1}\hat{X}'Y$). While both expressions are equivalent when the first stage is estimated by ordinary least squares, the values differ when a nonlinear relationship is estimated, and only the latter expression remains a consistent estimator for β (Angrist and Pischke, 2009).

Recent work highlights the weakness of some conclusions based on two stage least squares estimation when used in practice. For example, Young (2019) points out that, "In published papers, statistically significant IV results generally depend upon only one or two observations or clusters, excluded instruments often appear to be irrelevant, there is little statistical evidence that OLS is biased, and IV confidence intervals almost always include OLS point estimates." Dieterle and Snell (2014) also explore the sensitivity of model parameter estimates to changes in the first stage specification by allowing for the possibility of a quadratic relationship between instrument and endogenous regressor, as compared to constraining the relationship to a linear one. They observe that, "Across the fifteen papers [they] study here, [they] find evidence of significant nonlinearities in ten papers. Six of these ten studies have cases where the significant quadratic first stage is associated with a statistically significant difference in the 2SLS estimates of interest."

Based on the results we present in this paper, we do not find the above surprising. In particular, we show that if the true relationship in the first stage is quadratic, it is possible to draw a sample in which a linear first stage appears statistically significant, but which then produces a model parameter estimate that lies on the opposite side of the OLS estimate than does the true value, leading to incorrect inference. We also see that two stage least squares can be very sensitive to the sample drawn and can produce wide confidence intervals, as well as point estimates that are further from the truth than is the OLS. While not a panacea, the neural network based approach results in a smaller variance of the resulting estimator, which typically leads to a smaller mean squared error in parameter estimates. Overall we observe little harm from allowing for nonlinearity in

estimating the first stage, and much advantage.

The remainder of the paper proceeds as follows. Section 2 discusses nonlinear estimation methods and introduces neural networks. Sections 3 and 4 describe our simulation set-up and estimation process, respectively. Section 5 presents the results. Section 6 discusses extensions. Section 7 concludes.

2.2 Background

To some extent, the choice to use neural networks is an arbitrary one. Our main intent is to allow for potential nonlinearities in the first stage and then to calculate the impact on the estimates of interest in the second stage. This is not the first paper to suggest nonlinear IV. The theory has been developed previously (e.g. Newey et al., 1999; Newey and Powell, 2003; Blundell et al., 2007; Horowitz, 2011; Chen and Pouzo, 2012). Neither is this the first paper to suggest the use of neural networks in IV estimation – this has been done by Hartford et al. (2017).

The papers mentioned above focus on expanding the toolbox available to researchers and on describing the properties of the tools which they develop. Our aim is to bridge the gap between theory and applied work by providing an intuitive practical guideline for the extent to which estimates can suffer when we ignore potential nonlinearity in practice. We also identify situations where inference based on a linear method leads to a confident but incorrect conclusion. In applied work, we would encourage the estimation of the first stage both linearly and nonlinearly, to compare the resulting second stage estimates. If the two differ significantly, it is likely nonlinearity is present. Estimating the first stage using a neural network is a convenient way to allow for nonlinearity, and it appears to reasonably estimate a wide range of functional forms within a finite sample. Further work is necessary to fully establish the differences in finite sample performance of different nonlinear modeling methods, but we provide some indication below.

2.2.1 Flexibility of nonlinear models

When choosing a nonlinear modeling method, there are two aspects to consider. The first is flexibility – what class of functions can the method approximate well? This typically means choosing an appropriate set of basis functions, a linear combination of which then approximates the function being modeled. For instance, using a polynomial basis amounts

to estimating the Taylor expansion of the DGP over the support. Each additional term adds a further degree of flexibility, as the degree of the polynomial determines the maximal number of turns the polynomial can take.

If the researcher *a priori* knows that the DGP is periodic, using polynomials is not a good choice, as a polynomial approaches positive or negative infinity in the limit. Instead of a Taylor expansion, a Fourier series expansion may be more appropriate, where the basis functions are sinusoidal. Numerous other potential basis functions are possible and have been proposed, such as wavelets, splines, or Hermite polynomials. In a neural network with a single input and a single hidden layer, the basis functions are sigmoids[1], each of which can be linearly stretched or shifted.

Each additional node in a single hidden layer neural network corresponds to adding an $\alpha_1 \cdot \sigma(\alpha_2 + \alpha_3 x)$ term, where the x is the input (observation), $\sigma(\cdot)$ is the activation function, and the αs are parameters to be estimated. A bias (constant) term is also included. Adding further layers complicates the basis function reasoning, as each additional layer treats the output from the previous layer as inputs. This means that we end up with activation functions applied to activation functions, resulting in greater complexity.

All the methods discussed so far are universal approximators. In the limit, they can approximate any continuous function. In theory, it does not matter which basis we choose. As the number of terms included goes to infinity, any method will trace out any function to any preset level of precision. In practice, we use a finite number of terms, and it is easier to interpret a terser representation than one with numerous terms. For instance, it is easier to understand $\sin(x)$ than a summation of polynomials. The "correct" choice of basis functions – one that will approximate the DGP using the fewest terms – depends on foreknowledge of the specific DGP. Even if we choose the "correct" basis, we still need to ensure we include enough terms. The excluded terms are restricted to have a weight of 0, which may preclude reaching the desired level of precision in the approximation. An advantage of neural networks is that they make it easy to add a lot of complexity very quickly, especially with the addition of each new layer. As such, one is unlikely to specify insufficient model flexibility. This is useful, as in practice we do not know the true shape of the DGP we aim to estimate. On the other hand, interpreting the resulting function may be difficult, so this approach has mainly been used when the focus is on creating a prediction rather than understanding the underlying process.

[1]The basis function is whichever activation function is chosen. Typical choices given the estimation technique used include sigmoids or the rectifier.

2.2.2 Estimation of nonlinear models

The second aspect to consider when choosing a nonlinear modeling method is estimation. Some unknown values of the parameters in the model will bring us closest to the true DGP function. We can estimate the same model using standard OLS, a regularized version such as LASSO, maximum likelihood methods, or, in the case of neural networks, stochastic gradient descent (Robbins and Monro, 1951; Bottou, 2010). Each approach will in practice produce different estimates for each parameter, and even the same method may be computed slightly differently by different software or by using different numerical approaches. A further complication is that the sampled values from the DGP that we have available may be noisy measures, and we may not have very many measurements.

The closed-form solution for parameter estimates given by OLS may be most familiar, but all of the estimation methods described above are based on the minimization of a loss function. (For maximum likelihood, we can minimize the negative of the likelihood function.) Typically, the loss function is the sum of squared residuals, as it is in OLS, but it can take other forms, such as the sum of the absolute values of the residuals. The design of loss functions is its own area of research (e.g. Masnadi-shirazi and Vasconcelos, 2008). Regularizing terms are added to the loss function with a weight, pushing the loss upwards when parameter values are set to nonzero values, forcing a cost-benefit trade-off on the estimation. Calculating the derivatives with respect to each parameter and then solving the resulting system of equations is not always possible[2], either analytically or because computing the values directly is very costly, so approximations become necessary.

Stochastic gradient descent is one such approach. The basic idea is to take "steps" through the space spanned by the parameters until the steps taken by the algorithm no longer decrease the value of the loss function. The algorithm is initialized at a random point – a random value for each parameter – and evaluates the gradient at that point with respect to each parameter. A random subset of parameters is then chosen and the average gradient across that subset is calculated. This determines the direction in which the algorithm will step. The initial size of the step must be specified, but after that it is possible to adjust it dynamically. After taking the step, the parameters are set to their new values and the new value of the loss function is evaluated. The algorithm continues until the value of the loss function no longer decreases sufficiently with each new step, at which point it concludes that the minimum value has been reached. Bishop (2006) contains a more

[2]Recent work by Dyer and Gur-Ari (2019) has shown that it is possible to calculate a closed-form solution for some terms that govern neural network training, but this is currently not standard practice.

detailed discussion.

When estimating impacts in an instrumental variables framework, the second-stage coefficients will be affected if the first stage $\hat{X}$ predictions change. Constraining the first stage relationship to be linear limits the precision with which it is possible to approximate the true relationship. The further the true DGP is from linearity, the more constraining is this bound. Even a perfect estimate of the coefficients of a linear first stage model will then mispredict the way in which X varies with Z, the instrument. This, in turn, will affect the inference of how Y is affected by changes in X. At best, it will make the estimate less precise; at worst, it will lead to erroneous conclusions about the nature of that relationship.

Neural networks provide an easy way for a researcher to maximize their agnosticism. Allowing a great deal of flexibility ensures that the relationship between X and Z can potentially be estimated with great precision. If the neural network overfits and estimates not only the variation in X due to Z but also the variation due to unmodeled noise, this would bring the second stage estimates towards the OLS estimates, as $\hat{X}$ would be nearly the same as X. So long as the overfitting is not biased across the support, for example due to heteroskedasticity, we should be in no danger of "inventing" a relationship that is not truly there.

2.2.3 Comparing neural network with LASSO

To evaluate how well a neural network estimates a nonlinear DGP compared to a more standard approach, we compare its finite sample performance to LASSO. A typical starting point if one suspects nonlinearity is present in the relationship between two variables y and x is to add powers of x to the model, a polynomial basis. Regularization then "weeds out" the irrelevant powers of x during estimation by setting the weight on those terms to 0.

Because we intend to utilize relatively simple DGPs in the simulation, we set up a relatively small neural network with a single hidden layer of 5 nodes. We keep the same configuration for the simulations reported later in the paper in order to keep all results comparable. While it is possible to quickly scale the complexity, we merely wanted to introduce the possibility of nonlinearity into the estimation, so we erred on the side of too little flexibility. In typical applications, neural networks will be set up with hundreds of nodes and numerous hidden layers.

With a single hidden layer, each node is associated with 3 free parameters, plus the model allows for a bias, so the total number of parameters in the model set up is 16. We use the hyperbolic tangent activation function and estimate the model using stochastic gradient descent. Both of these are common choices, although other options are available. The estimation method and activation function are typically chosen based on resulting performance. In our case, the intention was to keep results comparable using some default settings across a variety of scenarios, so we did not try to improve on this combination. The estimation method includes regularization, so not all parameters may end up being used. To keep the comparison fair, we include a constant and fifteen powers of x in the LASSO estimation.

In each case, we aim to estimate $f(x)$, where $y = f(x) + \epsilon$. We compare how closely the predicted values of $\hat{y}$ match the true relationship between y and x by comparing the mean squared error of each set of estimates, that is, $(1/n)\sum_{i=1}^{n}(f(x_i) - \hat{y}_i)^2$. Here, n is the sample size, which we set to $n = 10,000$. We also consider two possible $f(\cdot)$ functions, x^2 and $\cos x$ – in the first case, the "correct" basis has been chosen for the LASSO, and in the second case, it is "incorrect". Both x and ϵ are assumed to be repeated i.i.d. draws from two different zero-mean normal distributions. The variance of x is set to 15. This means we expect to observe values of $f(x)$ approximately between 0 and 135 under the first DGP and between -1 and 1 in the second DGP. We set the variance of ϵ to either be low, about 20% of the difference between max and min $f(x)$, or high, about 70% of the difference. In the first case the corresponding values are 27 and 94.5, and in the second case it is .4 and 1.4.

In the quadratic case, where we chose the "correct" basis, LASSO unambiguously outperforms the neural network. Its MSE is 813.55 and 916.51 in the low and high noise case, respectively. The corresponding values for the neural network are 578491.99 and 1052041.37. In this case it appears that the neural network we set up was too small and ran out of flexibility, as it does not fit well in the sparser parts of the support, where x is largest[3].

In the cosine case, however, the neural network performs better. It does still appear to run out of degrees of freedom and is not able to fit well at the edges of the support as data becomes sparser. Overall, however, it still performs much better than the LASSO, which appears not able to fully make use of the flexibility it has. In the low noise case, the MSE for LASSO is 4270.71 and 970.18 for the neural net. In the high noise case, the values are

[3]Changing the activation function to the rectifier or adding a greater number of nodes to the hidden layer does improve the fit.

4271.73 and 1153.58.

Figure 2.1 shows the true relationship and both predictions for each DGP in the low noise case. Even with a small single hidden layer, a neural network appears better able to conform to a variety of shapes. Its estimation process may be more efficient at finding the best set of parameters as compared to the LASSO.

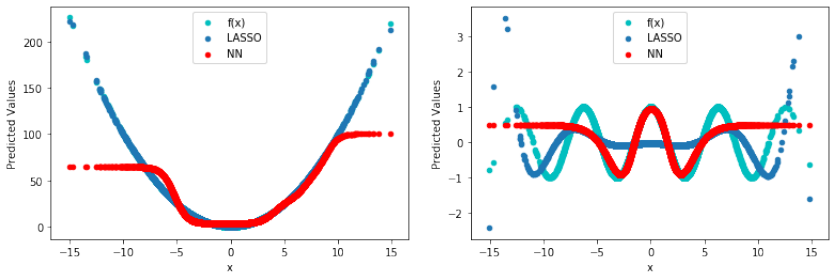


Figure 2.1: LASSO vs NN performance in the low noise case (noise not shown)

2.3 Simulating data

To determine the relative performance of the neural network (NN) approach compared to standard two stage least squares (2SLS), we set up a data generating process to create synthetic data. For each experiment, we repeatedly generate data samples using the DGP, as we would if we could repeatedly draw samples from a population for analysis. This allows us to empirically trace out the "true" distribution of parameter estimates for a sample of a given size, from which we usually observe only one draw in real-world applications.

2.3.1 Basic set-up

The process used to generate the data is the following:

$$\begin{aligned} &\textit{Model: } Y = X\beta + \epsilon \\ &\textit{First stage: } X = f(Z, \gamma) + \nu \end{aligned} \tag{2.1}$$

Table 2.1 details all the versions of the $f(\cdot)$ function considered. The aim is to estimate the value of β as closely and accurately as possible in each case. X is endogenous because the error terms in the model and first stage are set such that $\text{Cov}(\epsilon, \nu) \neq 0$. The $f(\cdot)$ function values are generated independently of the error terms. We report the OLS estimates, which show the bias that results from the endogeneity, and the $\hat{\beta}$ estimates resulting from

Model	
$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$	
First stage DGPs	
$x_i = f(z_i, \gamma) + \nu_i$	
Linear	$x_i = \gamma_0 + \gamma_1 z_i + \nu_i$
Logistic cdf	$x_i = 200\left(\frac{1}{1+\exp(\frac{-(\gamma_0+\gamma_1 z_i)}{0.15})}\right) + \nu_i$
Quadratic	$x_i = (\gamma_0 + \gamma_1 z_i)^2 + \nu_i$
Sinusoid	$x_i = 90\cos(0.2(\gamma_0 + \gamma_1 z_i)) + \nu_i$
Sinusoid2	$x_i = 90\sin(0.2(\gamma_0 + \gamma_1 z_i)) + \nu_i$
Probit	$x_i = \begin{cases} 1 & x_i^* = f(z_i, \gamma) + \nu_i > 0 \\ 0 & x_i^* = f(z_i, \gamma) + \nu_i \leq 0 \end{cases}$ $f(z_i, \gamma)$ takes on the same functional forms as above, except the logistic and quadratic are shifted downwards by 100 and by the value of σ_ν, respectively. $\nu_i \sim N(0, \sigma_\nu^2)$
Unless otherwise noted, the following parameter values are used: $\beta_0 = 12$, $\beta_1 = 1$, $\gamma_0 = 0$, $\gamma_1 = -3$, $\sigma_z = \sqrt{10}$, $\sigma_\epsilon = 40$, $\sigma_\nu = 200$, $\rho = -0.8$	

Table 2.1: Functional forms used to generate different kinds of nonlinear relationships between the instrument and endogenous regressor.

estimating the first stage relationship either using the standard linear method or a neural network.

The baseline case is linear, where $f(Z, \gamma)$ is simply $Z\gamma$. Z contains any exogenous regressors from the model and the instrument, and γ are the associated parameters. For every version of first stage DGP considered, any exogenous X variables and instrument are drawn from a joint normal distribution centered at the origin, with the covariance matrix chosen in such a way as to ensure a bias resulting from endogeneity is present and sufficiently large. The two error terms, ϵ and ν, are drawn from a different joint normal distribution to ensure independence, so that in expectation neither is correlated with the exogenous variables. We set the correlation ρ between ϵ and ν as well as the corresponding standard deviations, σ_ϵ and σ_ν, which determines the covariance of the joint normal mean 0 distribution. Once the exogenous variable data has been generated, the endogenous variables y_i and x_i are formed as shown in equation 2.1 using a specific choice of $f(\cdot)$.

2.3.2 Generating bias

To elaborate further on the choice of parameter values for the simulation, recall the source of the bias in the OLS estimates. Consider the simplest linear model with a single endoge-

nous regressor x_i and corresponding instrument z_i:

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_i + \epsilon_i, \text{ where} \\ x_i &= \gamma_0 + \gamma_1 z_i + \nu_i \text{ and } \text{Cov}(\nu_i, \epsilon_i) \neq 0 \end{aligned} \tag{2.2}$$

The object that $\hat{\beta}_{1,OLS}$ estimates will be $\frac{\text{Cov}(x_i,y_i)}{\text{Var}(x_i)}$. This is the true value of β_1 plus $\frac{\text{Cov}(x_i,\epsilon_i)}{\text{Var}(x_i)}$. Plugging in the expression for x_i, we can write the bias in $\hat{\beta}_1$ in terms of parameters and exogenous variables as the following:

$$bias = \frac{\rho\sigma_\epsilon\sigma_\nu}{\gamma_1^2\sigma_z^2 + \sigma_\nu^2} \tag{2.3}$$

All of the values in the above expression are parameters of the simulated DGP that we can set to ensure a non-zero bias. As the model becomes more complex, the exact expression for the bias becomes more difficult to work out. However, it can be shown that when there is one additional exogenous variable, the bias term of the parameter estimate on the endogenous variable becomes the following, where it is again possible to set all the values:

$$bias = \frac{\sigma_{x_1}^2\rho\sigma_\epsilon\sigma_\nu}{\sigma_{x_1}^2(\gamma_0^2 + \gamma_1^2\sigma_{x_1}^2 + \gamma_2^2\sigma_w^2 + 2\gamma_1\gamma_2\sigma_{x_1,w} + \sigma_\nu^2) - (\gamma_1\sigma_{x_1}^2 + \gamma_2\sigma_{x_1,w})^2} \tag{2.4}$$

In the above, x_1 is the additional exogenous variable, x_2 is the endogenous variable which is modeled linearly as $\gamma_0+\gamma_1 x_1+\gamma_2 w$, and w is the instrument. $\sigma_{x_1,w}$ represents $\text{Cov}(x_{1i}, w_i)$, and the σ^2 terms represent variances.

In his paper, Young (2019) notes that there is little statistical evidence of bias in OLS estimates. We found that in setting parameter values, it is easier to generate smaller sizes of bias, as the parameters counteract each other's effects or simply drive the bias towards zero. An interesting question to consider in future research is what fraction of the parameter space actually generates a large enough bias to be statistically detectable or economically relevant, and in which cases the bias is most likely to be observed in practice.

2.4 Estimation

Once we have generated the data, we proceed to estimate the parameters of interest (β) using the 2SLS and NN approaches. Each experiment consists of 2000 samples drawn from the population DGP and analyzed. For a single iteration of an experiment, we draw a

data sample of a pre-specified size and then construct $\hat{\beta}_{2SLS}$ and $\hat{\beta}_{NN}$, our two estimates of β. (We also calculate $\hat{\beta}_{OLS}$ as a benchmark.) The second stage is estimated identically in both cases, using the predicted values of the endogenous variable as instruments: $\hat{\beta}_j = (\hat{X}_j'X)^{-1}(\hat{X}_j'Y)$, $j \in \{\text{2SLS, NN}\}$. Only the estimation of the first stage differs.

For $\hat{\beta}_{2SLS}$, the first stage is calculated in the usual way, linearly using OLS by regressing the endogenous regressor on any exogenous regressors and the instrument. $\hat{X}_{2SLS}$ is then simply $Z\hat{\gamma}_{OLS}$.

We estimate the $\hat{X}_{NN}$ in MATLAB, using the Deep Learning Toolbox. The input to the neural network is Z, the matrix of observations of any exogenous regressors plus instrument, and the corresponding targets are the vector of values of the endogenous variable. We use a single hidden layer, with Bayesian regularization backpropagation as the estimation algorithm. The size of the hidden layer can be easily adjusted to allow for a more or less complex model; unless otherwise noted, we set the size of the hidden layer to 5 for all experiments.

In our experiments, one iteration thus produces one $\hat{\beta}_{2SLS}$ and one $\hat{\beta}_{NN}$. We can see how these vary for a given sample size when we draw multiple data samples from the population; a histogram shows the distribution of each estimator across all 2000 iterations. The true variance of each $\hat{\beta}$ is then estimated using the variance of this empirical distribution.

2.5 Results

Standard 2SLS is a robust method, and generally performs well, so long as there is an overall slope in the first stage for it to uncover. Even so, the neural network approach typically produces an estimator with a lower mean squared error. That is, although it can in some cases be slightly more biased in the finite sample than 2SLS, this is more than compensated for by the reduction in the variance of the parameter estimates. In certain cases, estimating the first stage nonlinearly results in both less bias as well as a lower variance.

An instrument may appear weak when it is not – the linear estimate has a slope close to 0, but a strong nonlinear relationship exists[4]. That is usually relatively harmless and necessitates looking for a different instrument, if using standard 2SLS. However, it is also

[4]We know that here because we generate the data based on a specific nonlinear relationship. Determining whether this is true in an arbitrary dataset is more difficult and the issue is discussed further in section 2.6.

possible that in the linearly estimated first stage, the instrument appears to be statistically significant, but this significance is the consequence of drawing a sample in which there is an overall slope, not a result of approximating the DGP correctly. In this case, the erroneous linear estimate of the DGP is not simply an oversimplification of the process, but it is an incorrect estimate of the average slope. This results in misleading estimates of $\hat{\beta}$, which can yield incorrect qualitative conclusions, i.e. that OLS over- rather than under-estimates the true effect.

For an instrumental variables approach to work, it must be the case that there be a relationship between the instrument(s) and the endogenous regressor, the stronger, the better. This is usually verified using an F statistic from the first stage regression (although, as described above, this can be misleading). Importantly, what is actually needed is a strong relationship between the instrument and the *exogenous* portion of the endogenous regressor, as the aim is not to match x using z as closely as possible, but to extract the relevant signal from the endogenous noise; in other words, to find the relevant $f(Z, \gamma)$. Failure to disentangle the signal from the noise biases an instrumental variable estimate towards the OLS estimate (Mullainathan and Spiess, 2017).

Correspondingly, in the tables and figures below, we report $\overline{MSE(f)}$, a modification of the standard sum of squared residuals measure. This measures how closely each approach predicts the exogenous portion of the endogenous variable. More concretely, for a single experiment $MSE(f) = \sum_{i=1}^{n}(\hat{X}_i - f_i)^2/n$, where n is the sample size, and the reported measure $\overline{MSE(f)}$ is the average across all experiments. The lower this measure, the closer are the estimated values to the true relationship between x and z. Certainly this is not possible to observe when one does not know the true DGP, but seeing the relative performance of a linear versus nonlinear first stage fit in our experiments provides a useful benchmark.

2.5.1 Continuous endogenous variable

Model: $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$
First stage: $x_i = f(z_i, \gamma_0, \gamma_1) + \nu_i$, $\text{Cov}(\epsilon_i, \nu_i) \neq 0$

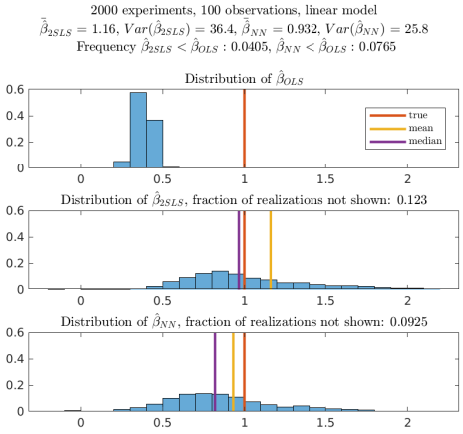
Table 2.1 lists the functional forms of $f(\cdot)$ considered. As one might expect, 2SLS holds the biggest advantage when the first stage DGP is truly linear. Even so, it does not win on every metric, and the neural network approach is not far behind. We look at several cases, illustrated in figures 2.2 and 2.3.

Figure 2.2 considers the relatively less noisy case, where we set $\sigma_\nu = 50$. Figure 2.2(f) gives some intuition of what the noisy (X) and de-noised ($f(Z, \gamma)$) data look like. In the figure we show an example of the first stage predictions from a single experiment iteration with 500 observations, chosen randomly, based on the data that was drawn. The x-axis measures the value of $Z\gamma$, and the y-axis measures the value of the corresponding estimates. Since the DGP is linear in this case, the true signal corresponds to the 45° line through the origin. The scattered points are the actual values of X, and the other two data series are $\hat{X}_{2SLS}$ and $\hat{X}_{NN}$. From the image it is obvious that the linear method more closely approximates the signal in this case, and the neural network slightly overfits, picking up on more of the noise where data is sparser. We can see this is true on average across all 2000 experiments of sample size 500, as the $\overline{MSE(f)}_{2SLS}$ is smaller than the $\overline{MSE(f)}_{NN}$. As the first five rows of table 2.2 show, the fit of both models improves with larger sample sizes, although it remains true that $\overline{MSE(f)}_{2SLS} < \overline{MSE(f)}_{NN}$ at each sample size in the linear DGP case.

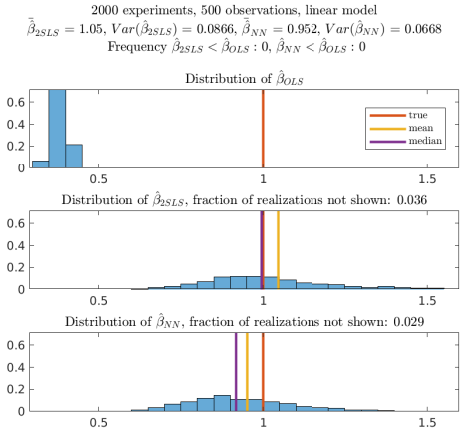
Understandably, the fit is worse when the first stage error has higher variance ($\sigma_\nu = 200$), as it becomes more difficult to separate the signal from the noise for either method. One can see the increase in the scatter of the data in figure 2.3(f). Table 2.2 shows that here too, 2SLS manages to achieve a better fit, and for both methods the fit improves with sample size.

In terms of estimating the parameter of interest, the slope on the endogenous variable, we show the distribution of $\hat{\beta}_{1,OLS}$, $\hat{\beta}_{1,2SLS}$, and $\hat{\beta}_{1,NN}$ achieved across 2000 experiments using datasets of various sizes. The intuition is once again that of repeated draws from a population; in reality we would typically only have access to one such draw, so the properties of the distribution from which this draw would be taken are important. In each figure 2.2(a) to 2.2(e) and 2.3(a) to 2.3(e), we report the means and variances of $\hat{\beta}_{1,2SLS}$ and $\hat{\beta}_{1,NN}$, based on the empirical distribution observed at different sample sizes, as well as the proportion of the time one would conclude erroneously that the bias in the OLS estimate is positive rather than negative. Such incorrect qualitative inference is concerning in applied work, as it can lead to unhelpful policy implications.

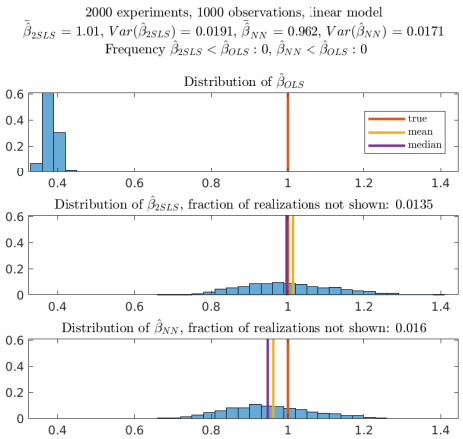
Every histogram of $\hat{\beta}_{2SLS}$ and $\hat{\beta}_{NN}$ is labeled with the fraction of realizations not shown. These are the automatically-determined outliers (using the MATLAB `isoutlier` function) that are omitted from the histogram in order to keep all histograms readable. The absence of these observations explains why a histogram's mean may appear in an unintuitive location, or potentially not appear at all within the limits of the x-axis shown.



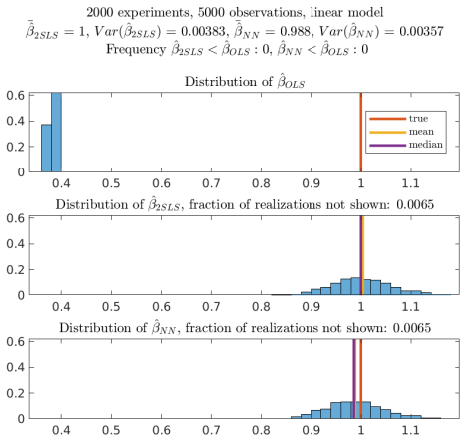
(a) Linear DGP, $\sigma_\nu = 50$, sample size = 100



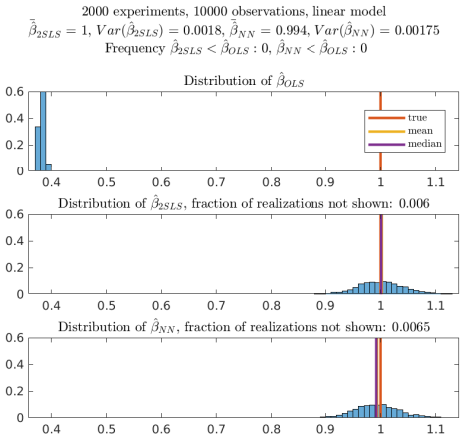
(b) Linear DGP, $\sigma_\nu = 50$, sample size = 500



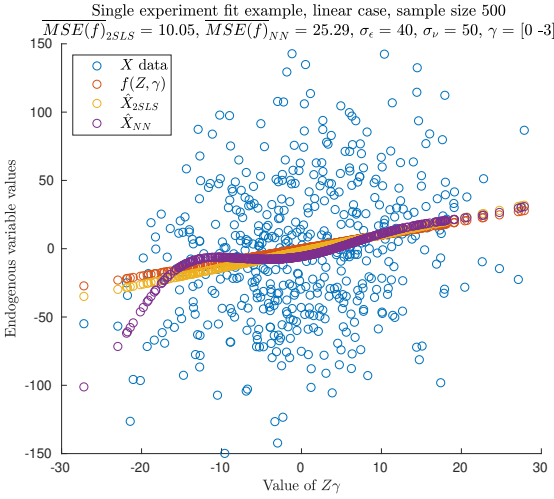
(c) Linear DGP, $\sigma_\nu = 50$, sample size = 1000



(d) Linear DGP, $\sigma_\nu = 50$, sample size = 5000

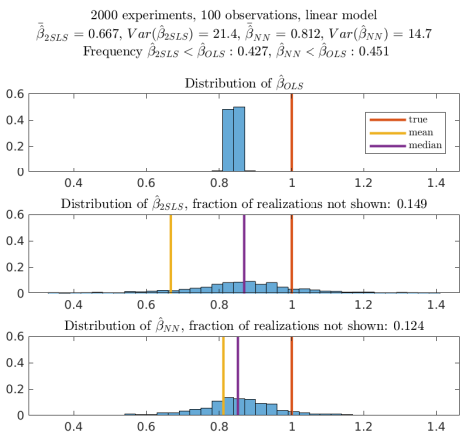


(e) Linear DGP, $\sigma_\nu = 50$, sample size = 10000

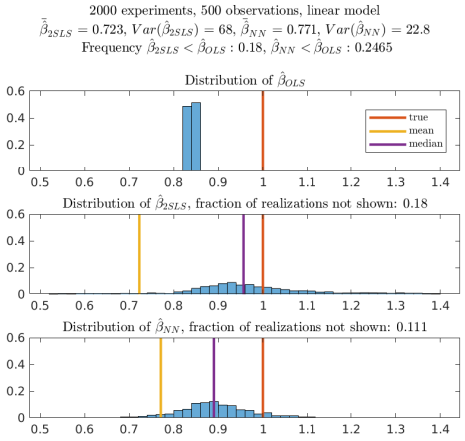


(f) Linear DGP, $\sigma_\nu = 50$, sample size = 500

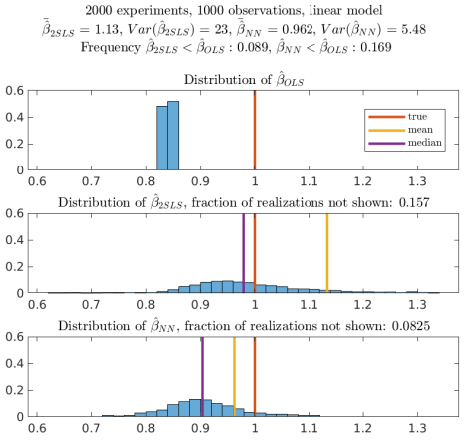
Figure 2.2: Results for linear first stage DGP at various sample sizes: low noise, $\sigma_\nu = 50$



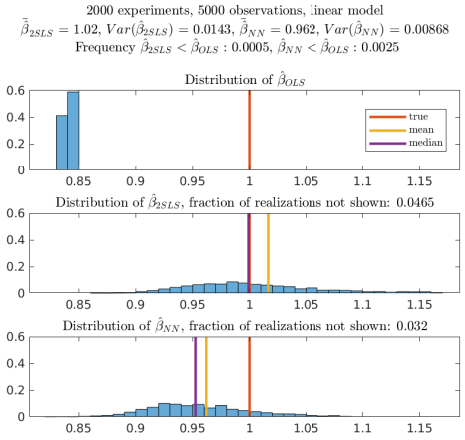
(a) Linear DGP, $\sigma_\nu = 200$, sample size = 100



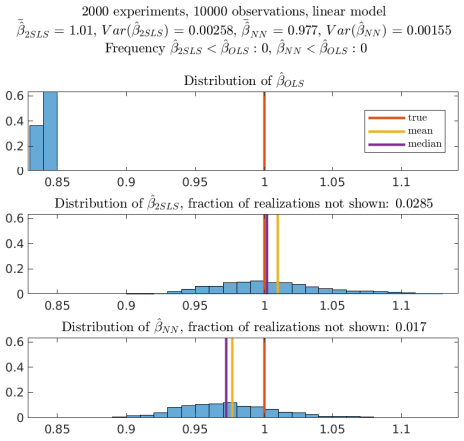
(b) Linear DGP, $\sigma_\nu = 200$, sample size = 500



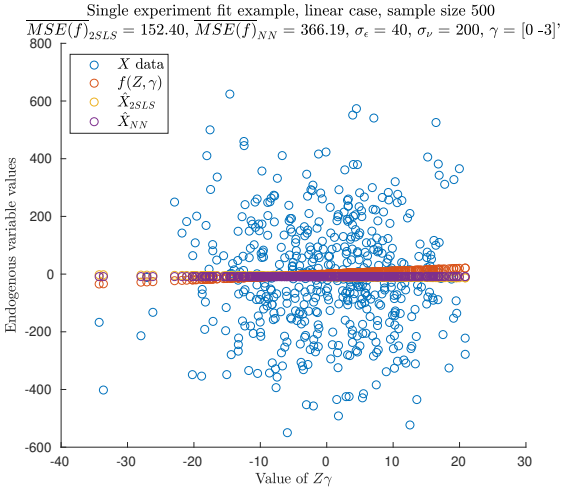
(c) Linear DGP, $\sigma_\nu = 200$, sample size = 1000



(d) Linear DGP, $\sigma_\nu = 200$, sample size = 5000



(e) Linear DGP, $\sigma_\nu = 200$, sample size = 10000



(f) Linear DGP, $\sigma_\nu = 200$, sample size = 500

Figure 2.3: Results for linear first stage DGP at various sample sizes: high noise, $\sigma_\nu = 200$

With the relatively high amount of noise we have introduced, we see that a sample size of 100 is insufficient to guarantee good performance with either method, especially with a noisier first stage. When $\sigma_\nu = 200$, $\hat{\beta}_{2SLS}$ leads us to the wrong qualitative conclusion less frequently than does $\hat{\beta}_{NN}$, but not by much, 42.7% of the time versus 45.1% of the time for the $\hat{\beta}_{NN}$. The numbers improve substantially when $\sigma_\nu = 50$, dropping to 4.05% and 7.65%, respectively. As the sample size increases, the performance of both methods improves. At every sample size, the researcher is more likely to draw the correct conclusion regarding the direction of the bias in OLS using the 2SLS approach, but the variance of $\hat{\beta}_{NN}$ is consistently lower. If one happens to draw a particularly unfortunate sample, one may get a much larger mis-estimate of β_1 using 2SLS than one would using the NN.

As mentioned earlier, when the fitted values $\hat{X}$ pick up noise as well as signal, we can expect the resulting estimate of β to be biased towards the OLS estimate. With a linear first stage DGP, the neural network overall tends to fit the signal worse than does the standard linear approach (table 2.2). It is true that in figures 2.2 and 2.3 both the mean and median of $\hat{\beta}_{NN}$ tend to be closer to the $\hat{\beta}_{OLS}$ distribution than the mean and median of $\hat{\beta}_{2SLS}$. In some cases this actually appears to be a positive, as the combination of some unfortunate small sample draws and high noise produces a mean $\hat{\beta}_{2SLS}$ estimate that is much lower than the OLS estimates, when in fact OLS underestimates the true parameter (see figures 2.3(a) and 2.3(b)).

We know that standard 2SLS estimates, though consistent, are biased in finite samples, so just the fact that the $\hat{\beta}_{NN}$ are biased is not enough to discredit the approach. Because the neural net does slightly overfit the linear DGP, the method produces a higher variance in $\hat{X}$ than does the 2SLS, leading to a lower variance of $\hat{\beta}_{NN}$. We can consider whether the trade-off is worthwhile by calculating the mean squared error based on the empirical distributions we generate, which approximate the true distributions. Table 2.3 reports the squared bias of the parameter estimates plus the variance calculated across the experiments, $(\bar{\hat{\beta}}_1 - \beta_1)^2 + \text{Var}(\hat{\beta}_1)$. Based on this measure, the slight increase in bias appears to be worth the corresponding decrease in variance, as the neural network approach consistently produces a lower mean squared error.

To more closely examine the trade-offs between using a neural network fit versus a linear fit when the DGP is linear, we run an experiment with 10,000 iterations, each drawing a sample of 100 data points (keeping $\sigma_\nu = 50$). A larger number of iterations should generate more precise estimates of the true distributions of the $\hat{\beta}$s. Because neither method can be expected to perform faultlessly with such a small sample size, we can observe on which

$\overline{MSE(f)}$ from first stage estimation			
$f(z_i, \gamma)$	Sample size	2SLS	NN
Linear (low noise, $\sigma_\nu = 50$)	100	51.45	102.75
	500	10.05	25.29
	1000	4.99	13.90
	5000	1.01	3.62
	10000	0.51	1.93
Linear (high noise, $\sigma_\nu = 200$)	100	792.34	1344.67
	500	152.40	366.19
	1000	79.73	222.61
	5000	16.65	59.10
	10000	8.05	30.82
Logistic	100	4113.58	3887.34
	500	3531.55	1632.40
	1000	3458.37	1032.67
	5000	3403.57	262.15
	10000	3393.78	118.05
Quadratic	100	16211.52	2571.67
	500	16101.16	585.50
	1000	16262.30	310.73
	5000	16168.81	72.61
	10000	16179.94	38.10
Sinusoid	100	4560.17	3057.95
	500	3966.06	702.23
	1000	3909.85	351.73
	5000	3846.82	85.36
	10000	3838.57	47.08
Sinusoid2	100	3930.76	2934.19
	500	3375.65	716.87
	1000	3317.93	374.71
	5000	3264.44	83.51
	10000	3257.60	44.12

Table 2.2: The fit achieved by 2SLS and NN in estimating the first stage. Lower values indicate a closer approximation of the $f(\cdot)$ function. When the first stage DGP is linear, 2SLS is able to pick up on the underlying signal better than the NN; however, the opposite is true in every other case.

Bias-variance trade-off: MSE($\hat{\beta}$)

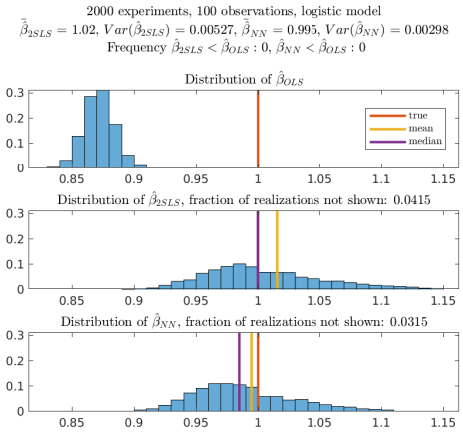
$f(z_i, \gamma)$	Sample size	2SLS	NN
Linear (low noise, $\sigma_\nu = 50$)	100	36.4684	25.8251
	500	0.0887	0.0691
	1000	0.0193	0.0186
	5000	0.0039	0.0037
	10000	0.0018	0.0018
Linear (high noise, $\sigma_\nu = 200$)	100	21.5109	14.7353
	500	68.0767	22.8524
	1000	23.0169	5.4814
	5000	0.0147	0.0101
	10000	0.0027	0.0021
Logistic	100	0.0057	0.0030
	500	0.000552	0.000398
	1000	0.000263	0.000196
	5000	0.0000531	0.0000360
	10000	0.0000258	0.0000173
Quadratic	100	11.6620	0.0239
	500	14.7094	0.0025
	1000	9.9449	0.0000989
	5000	128.1043	0.0000198
	10000	37.8169	0.00000938
Sinusoid	100	14.8204	1.6210
	500	9.3106	0.0830
	1000	10.1480	0.000476
	5000	16.842	0.0000843
	10000	17.7053	0.0000416
Sinusoid2	100	26.2067	1.0401
	500	0.0753	0.0012
	1000	0.0030	0.000416
	5000	0.000447	0.0000842
	10000	0.000213	0.0000406

Table 2.3: The MSEs of $\hat{\beta}_{2SLS}$ and $\hat{\beta}_{NN}$, calculated using $(\bar{\hat{\beta}}_1 - \beta_1)^2 + \text{Var}(\hat{\beta}_1)$. The values of the mean and variance are taken from the empirical distributions generated through simulation. In each of the above cases, $\text{MSE}(\hat{\beta}_{NN}) \leq \text{MSE}(\hat{\beta}_{2SLS})$.

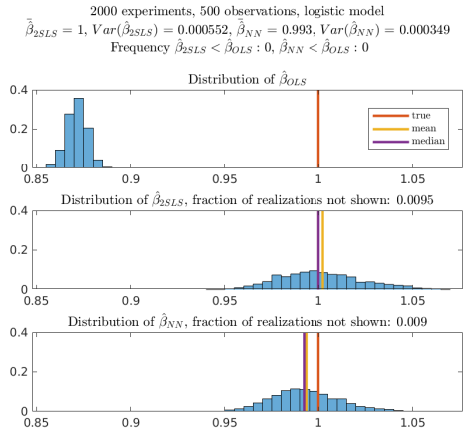
metrics each method does well and on which it falls short. In terms of the mean value, both methods overestimate the true parameter value, 1, but the $\bar{\hat{\beta}}_{NN}$ is closer, 1.25 as compared to 1.47. The median of the $\hat{\beta}_{2SLS}$ distribution is closer to 1, however, 0.9646 compared to 0.8168 for the NN. Both variances are high, but, predictably, the neural network estimator variance is lower. The sum of the squared bias and variance is again lower for the neural network approach, 1692.7 compared to 1866.8. The linear method results in less incorrect qualitative inference here, as $\hat{\beta}_{2SLS} < \hat{\beta}_{OLS}$ 4.08% of the time, as opposed to 7.34% of the time for the neural network. There are also more 2SLS estimates clustered around the true parameter value. The standard deviation of the empirical $\hat{\beta}_{OLS}$ distribution is 0.0492. If we look within two OLS standard deviations of the true β, we find 1789 2SLS estimates and 1649 NN estimates (of the 10,000). However, it is also true that we find more 2SLS estimates further away, which one expects, as that distribution has the higher variance. Beyond 10 OLS standard deviations away from the true β, we find 2594 2SLS estimates and 2381 NN estimates. It is worth noting that with this small sample size, only in 4505 instances did the $\hat{\gamma}_{1,OLS}$ achieve a $|t|$ value greater than 2. Of these, it was still the case that in 34 of these instances, $\hat{\beta}_{2SLS} < \hat{\beta}_{OLS}$, and the corresponding number is 67 for the neural network.

We next consider the logistic DGP. As can be seen in figure 2.4(f), the function we choose is quite nonlinear. In this case, the neural network is clearly better able to tease out the true f function from the noise than can the linear fit, as is evidenced by the $\overline{MSE(f)}$ values reported in table 2.2. Nonetheless, there is a clear positive slope for the linear method to pick up, and 2SLS performs well. The mean of the $\hat{\beta}_{2SLS}$ distribution approaches 1, the true value of β_1, more quickly than does the mean of the distribution of $\hat{\beta}_{NN}$, although the performance of the latter does not lag far behind in this regard. As before, the distribution of $\hat{\beta}_{NN}$ has a lower variance than does the standard 2SLS estimate at every sample size. This makes sense – as the neural network is able to fit the underlying signal more closely, the residuals $\hat{\nu}_i$ are closer to the actual realizations of ν_i. The estimated variance of the error is thus closer to the true value σ_ν^2, as opposed to overestimating it. Because the variance in the data is fixed, correspondingly the $\hat{x}_{i,NN}$ have a higher variance, leading to a lower variance of the $\hat{\beta}$. Table 2.3 shows the mean squared error of the second stage estimate for both methods. Once again, it appears that the decrease in variance generated by the neural network approach outweighs any increase in bias.

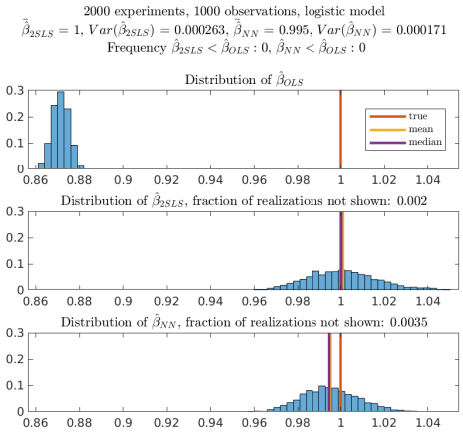
From a practical point of view, so long as there is a strong positive or negative relationship that 2SLS is able to pick up between the x and the z, regardless of which sample happens to be drawn, performance between the NN approach and the standard 2SLS approach is



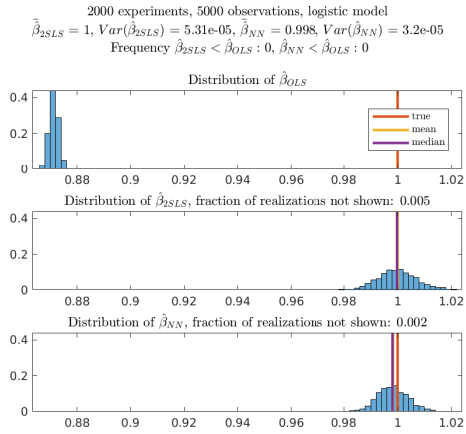
(a) Logistic DGP, $\sigma_\nu = 200$, sample size = 100



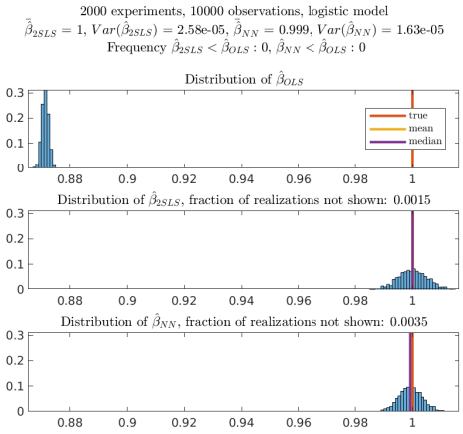
(b) Logistic DGP, $\sigma_\nu = 200$, sample size = 500



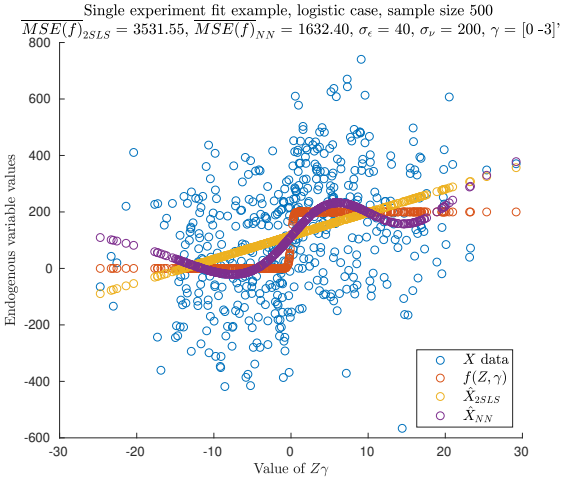
(c) Logistic DGP, $\sigma_\nu = 200$, sample size = 1000



(d) Logistic DGP, $\sigma_\nu = 200$, sample size = 5000

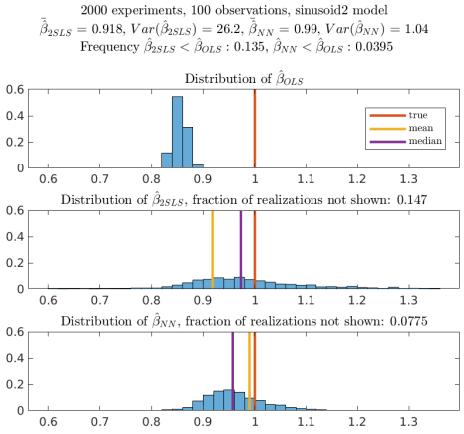


(e) Logistic DGP, $\sigma_\nu = 200$, sample size = 10000

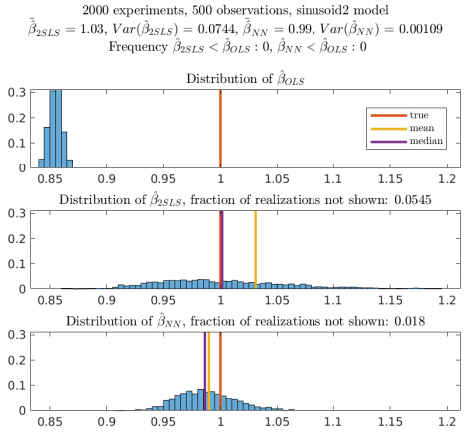


(f) Logistic DGP, $\sigma_\nu = 200$, sample size = 500

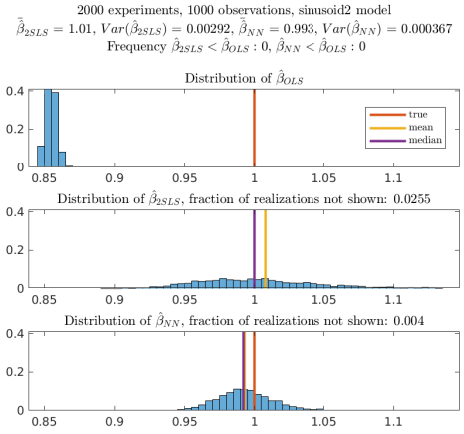
Figure 2.4: Results for logistic first stage DGP at various sample sizes



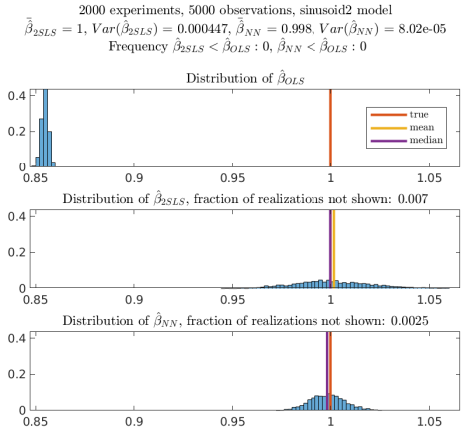
(a) Sinusoid2 DGP, $\sigma_\nu = 200$, sample size = 100



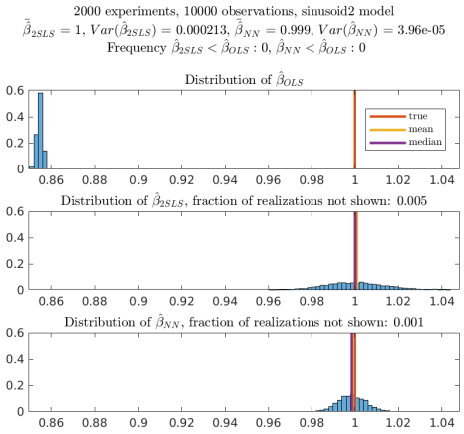
(b) Sinusoid2 DGP, $\sigma_\nu = 200$, sample size = 500



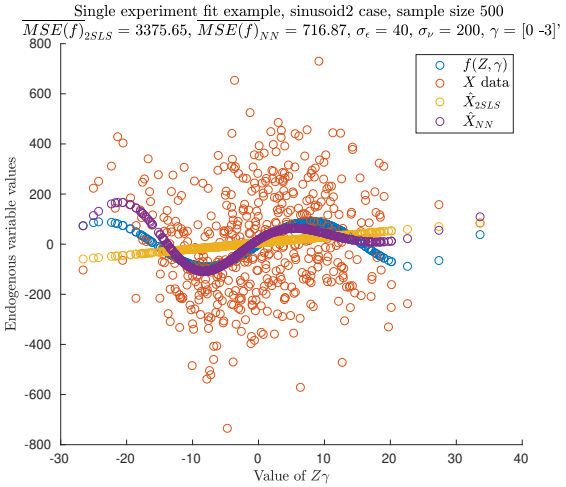
(c) Sinusoid2 DGP, $\sigma_\nu = 200$, sample size = 1000



(d) Sinusoid2 DGP, $\sigma_\nu = 200$, sample size = 5000



(e) Sinusoid2 DGP, $\sigma_\nu = 200$, sample size = 10000



(f) Sinusoid2 DGP, $\sigma_\nu = 200$, sample size = 500

Figure 2.5: Results for sinusoid2 (sine) first stage DGP at various sample sizes

fairly similar. Encouragingly, any potential overfitting by the neural network does not lead to misestimation of the parameters of interest, and in fact the neural network approach generates estimates with a lower mean squared error, even when there is evidence that the linear method fits the underlying (linear) signal more closely. We further test whether overfitting can be a problem by comparing performance of the two methods when there truly is no relationship between the instrument used and x. This is reported in the next section, 2.5.2, and the results are encouraging.

Of course, if the sample size is small and the average slope of the relationship between z and x is not so far from 0, the differences in performance of 2SLS and NN may become relevant. As an example, we present figure 2.5. The x here is a sine function of $\gamma_1 z$ (plus noise; see table 2.1 for details). As sine is an odd function and z is sampled from a mean 0 distribution, overall there tends to be a slope for the linear method to identify. However, this is weaker than in the linear or logistic cases, and one can see that with a smaller sample size, one is more likely to make an incorrect qualitative inference. The fraction of time $\hat{\beta}_{2SLS} < \hat{\beta}_{OLS}$ is 13.5%, whereas it is only approximately 4% when using the neural network first stage (sample size 100, figure 2.5(a)). Even with larger sample sizes, one can clearly see that the neural network first stage achieves a lower $\hat{\beta}$ variance without much risk of biased estimation; the NN estimates' mean squared error is once again lower (table 2.3). As in practical applications we typically only observe one sample, i.e. one draw from the DGP, drawing from the tighter distribution is more likely to result in a point estimate closer to the true parameter value.

We now discuss the type of case where 2SLS fails to perform and where the advantage of using a neural network becomes apparent. Figure 2.6 demonstrates what can occur if the $f(\cdot)$ is quadratic. We emphasize that it is unnecessary to guess at the exact nature of the relationship between z and x. One simply inputs the exogenous variables and/or instrument, and the neural network estimation derives the relationship from the data.

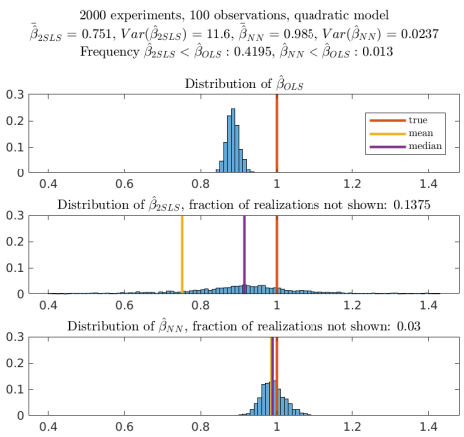
With a quadratic function, $x_i = (\gamma_1 z_i)^2 + \nu_i$, the average slope of the relationship between x and z tends to be approximately 0. Even though in truth there is a strong association between the two variables, z appears to be a very weak instrument when using the conventional method. This reflects in the performance of $\hat{\beta}_{2SLS}$, which produces untrustworthy estimates. Even at the largest sample size we use in our experiments, $\hat{\beta}_{2SLS} < \hat{\beta}_{OLS}$ over 40% of the time (figure 2.6(e)). Its variance and its bias are both high, and estimates of β_1 can be very far from the true parameter value, depending on the researcher's luck in drawing a data sample. The performance of $\hat{\beta}_{2SLS}$ does not improve in any way with

sample size. This can be seen in its mean squared error in table 2.3, which remains high and erratic, while the corresponding NN statistic becomes smaller and smaller. The neural network first stage fits the underlying signal much more closely (table 2.2) and produces a $\hat{\beta}$ distribution which becomes more and more centered on the true parameter value, becoming tighter as the sample size increases.

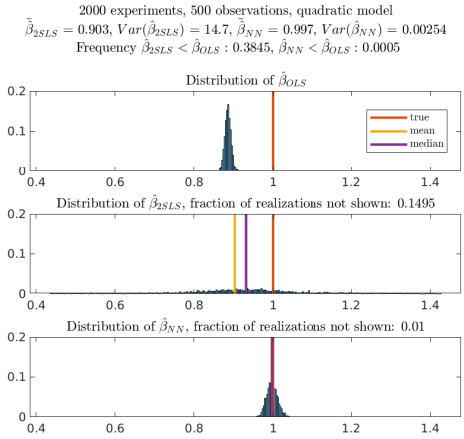
It is important to note that even when 2SLS is able to find a "strong" relationship in a particular data sample, this is no guarantee of a good estimate of β_1. For example, while in most samples the value of $\hat{\gamma}_1$ is not statistically significantly different from 0 under the quadratic first stage DGP, there are some samples in which the corresponding t-value is large enough to reject $H_0 : \gamma_1 = 0$. The instrument appears to be strong enough. We look at the $\hat{\beta}_{2SLS}$ that result from data samples in which $|t| > 2$. In one experiment of 2000 iterations, 226 samples meet this criterion. Of these, 26% produce a $\hat{\beta}_{2SLS}$ which is smaller than the corresponding $\hat{\beta}_{OLS}$, which could lead the researcher to the incorrect conclusion that $\hat{\beta}_{OLS}$ overestimates, rather than underestimates, the true effect β_1. In comparison, of these 226 cases, $\hat{\beta}_{NN}$ is never smaller than the corresponding $\hat{\beta}_{OLS}$, and even though the instrument appears "strong", in these samples the neural network approach produces estimates both less biased and more precise than does the 2SLS.

If one attempts to use an instrument which one expects to fulfil the validity and relevance requirements, perhaps relying on economic theory, it may be possible that the relationship between the instrument and the endogenous regressor is not simply linear, and perhaps the estimate of the average slope is not statistically significant. It is then worth exploring this possibility by estimating the first stage using a neural network. If there truly is no relationship, the neural network should be unable to find one, just as the linear method would fail. However, if there is a nonlinear relationship between the two variables, this may improve the resulting estimates and obviate the need to search for a different instrument.

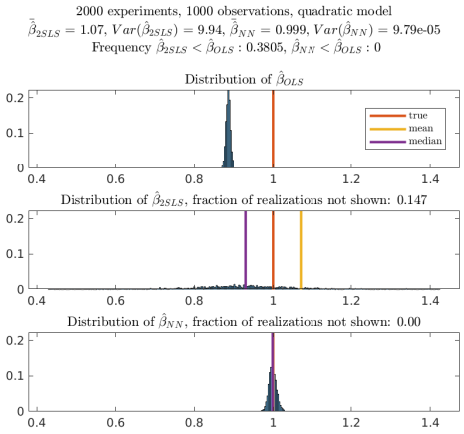
One can imagine any number of possible examples; we present one more. In figure 2.7, the first stage DGP is a periodic relationship, constructed (or could be sampled) in such a way that there is no overall positive or negative slope for a linear fit to pick up. Once again, $\hat{\beta}_{2SLS}$ suffers from high variance and bias, leading to $\hat{\beta}_{2SLS} < \hat{\beta}_{OLS}$ approximately half the time, at any sample size (so, is very much biased towards the OLS, as we would expect). The neural network does a better job of picking up on the underlying relationship between z and x, and produces estimates of β_1 which are both closer to the truth and more precisely estimated.



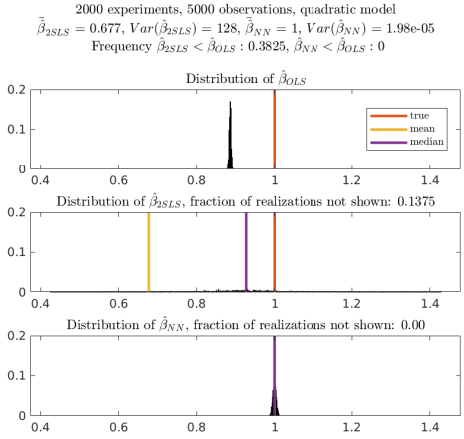
(a) Quadratic DGP, $\sigma_\nu = 200$, sample size = 100



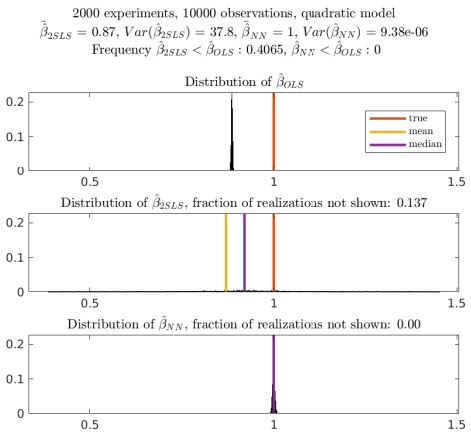
(b) Quadratic DGP, $\sigma_\nu = 200$, sample size = 500



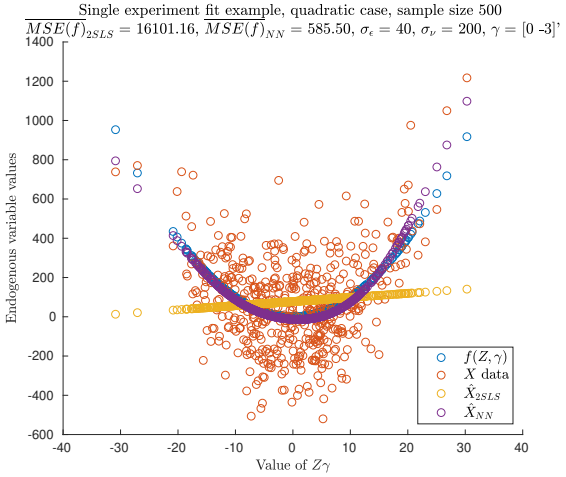
(c) Quadratic DGP, $\sigma_\nu = 200$, sample size = 1000



(d) Quadratic DGP, $\sigma_\nu = 200$, sample size = 5000

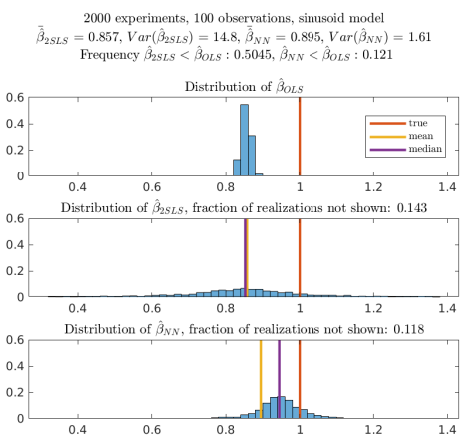


(e) Quadratic DGP, $\sigma_\nu = 200$, sample size = 10000

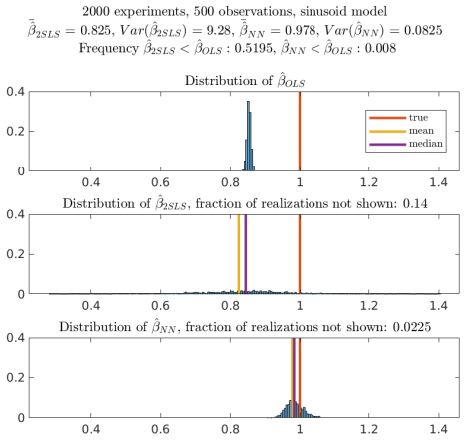


(f) Quadratic DGP, $\sigma_\nu = 200$, sample size = 500

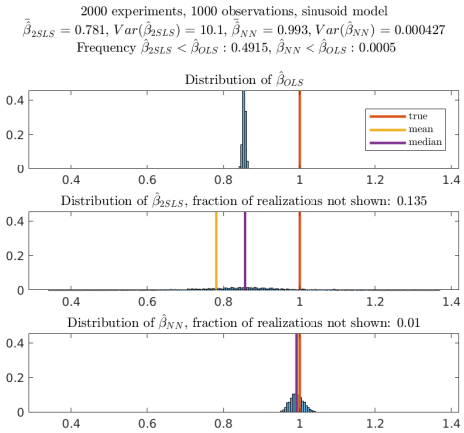
Figure 2.6: Results for quadratic first stage DGP at various sample sizes



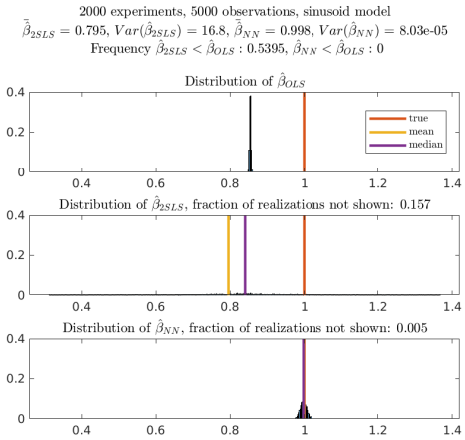
(a) Sinusoid DGP, $\sigma_\nu = 200$, sample size = 100



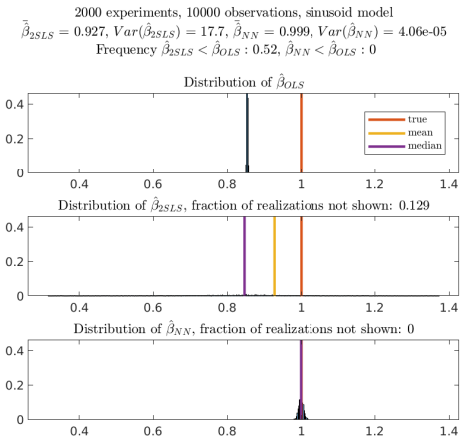
(b) Sinusoid DGP, $\sigma_\nu = 200$, sample size = 500



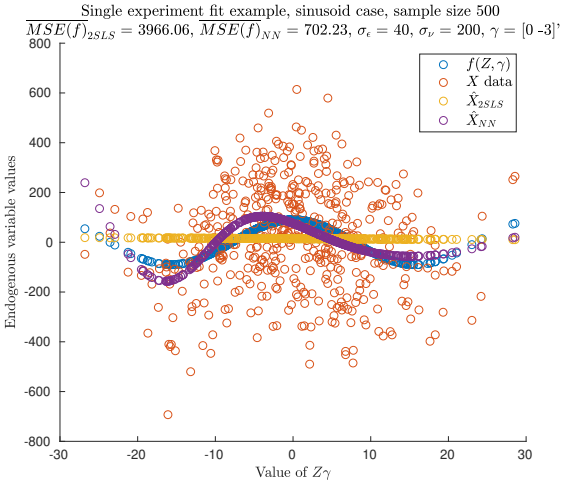
(c) Sinusoid DGP, $\sigma_\nu = 200$, sample size = 1000



(d) Sinusoid DGP, $\sigma_\nu = 200$, sample size = 5000



(e) Sinusoid DGP, $\sigma_\nu = 200$, sample size = 10000



(f) Sinusoid DGP, $\sigma_\nu = 200$, sample size = 500

Figure 2.7: Results for sinusoid (cosine) first stage DGP at various sample sizes

2.5.2 Weak instrument

A researcher may take comfort in knowing that if no relationship exists between x and the chosen z, there should be a statistically insignificant linear slope between the two. Earlier we demonstrated that even in the absence of a statistically significant linear slope, a strong *nonlinear* relationship may well exist, and it is clear that in those constructed examples the neural network does pick up on the actual truth. Could it be the case that a neural network would find what seems to be a similarly strong relationship even when none truly exists? Would it overfit to the noise and produce estimates of $\hat{\beta}$ that appear reasonable, even though they are incorrect?

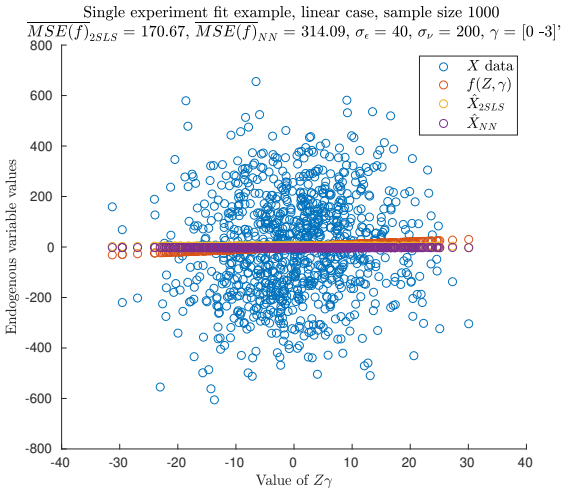
Figure 2.8 shows this to be unlikely, and in fact the lower variance of the $\hat{\beta}_{NN}$ estimates once again turns out to be an advantage. We touched earlier on what happens in the linear DGP when the signal is obscured by more or less noise ($\sigma_\nu = 50$ or 200), making the instrument relatively weaker or stronger. With a weaker instrument and smaller sample size, both the 2SLS and the NN estimates are biased towards the OLS. In the completely weak instrument case which we now explore (no relationship between the instrument used and x), we thus expect that the $\hat{\beta}$s should be estimating the same value as $\hat{\beta}_{OLS}$. This is indeed what happens, but the $\bar{\hat{\beta}}_{NN}$ is closer to $\bar{\hat{\beta}}_{OLS}$ in both of the cases considered, and its variance is much lower in both cases.

2.5.3 Dummy endogenous variable

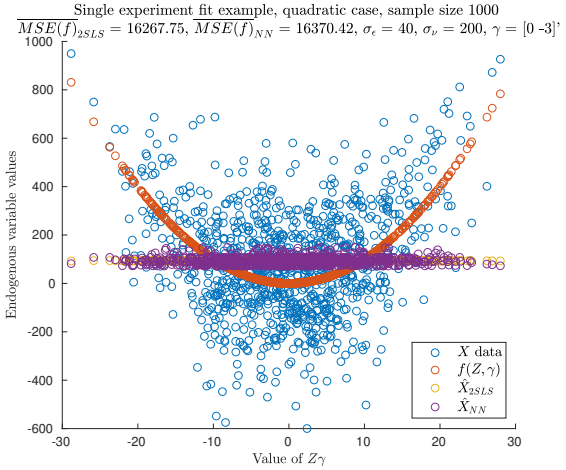
The most obvious nonlinearity one is likely to encounter in practice is a dummy endogenous variable. If the values of the instrument are drawn from a continuous distribution, one may be tempted to estimate the relationship using logit or probit. So long as the resulting $\hat{X}$ is used as an instrument in the second stage, this can be done, and may improve the efficiency of the estimates as compared to 2SLS. The estimation can also be carried out using a neural network, with similar benefits.

We model the dummy variable x using a latent variable model. Table 2.1 shows all of the DGPs considered. The basic approach is to model the value of x as depending on the value of an (unobserved in practice) x^*, which in turn is related to the instrument z through some function $f(\cdot)$, plus an error ν, which is distributed normally. We thus refer to this set of experiments as "probit".

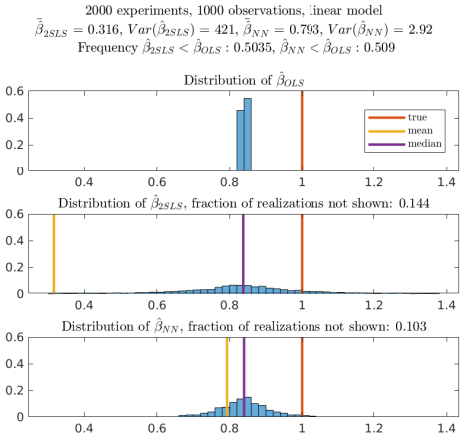
The simplest case is when the relationship is linear. Suppose z is an individual's age, $f(z, \gamma)$



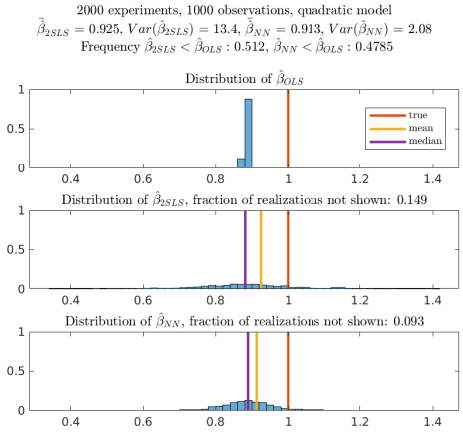
(a) Linear DGP, noise used as instrument, $\sigma_\nu = 200$



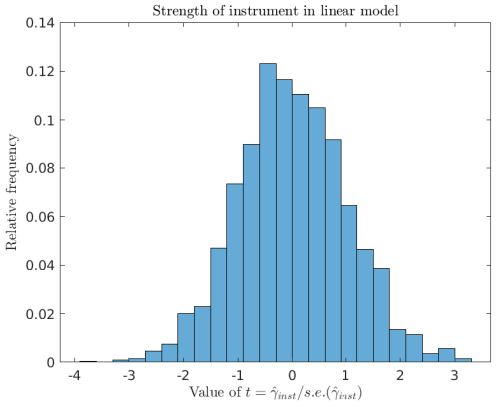
(b) Quadratic DGP, noise used as instrument, $\sigma_\nu = 200$



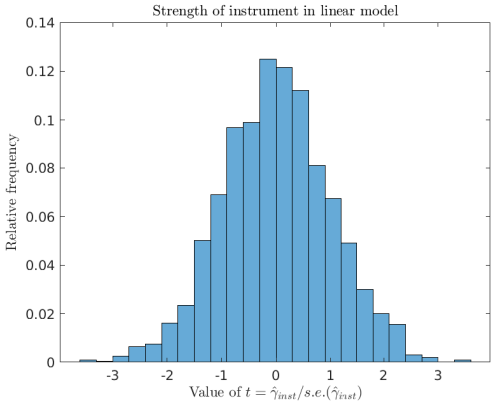
(c) Linear DGP, noise used as instrument, $\sigma_\nu = 200$



(d) Quadratic DGP, noise used as instrument, $\sigma_\nu = 200$



(e) Linear DGP, noise used as instrument, t values



(f) Quadratic DGP, noise used as instrument, t values

Figure 2.8: Weak instrument performance. Both approaches similarly fail to perform when the instrument used to form $\hat{x}_i$ in fact has no relation to the x_i (which is still formed using $f(z_i, \gamma) + \nu_i$, so z_i is the instrument that should be used). However, the $\bar{\hat{\beta}}_{NN}$ falls closer to $\bar{\hat{\beta}}_{OLS}$ (as the parameter estimates should, in this case) and has a lower variance in both examples.

is the true value gained for someone of that age from being enrolled in grade school, and ν is any individual uncertainty around that measure. We can then think of x^* as the value the individual (or her parents) thinks she derives from grade school attendance, and observe x, the binary decision of whether or not to enroll. In this example, one can suppose that at a young age, the value of school attendance is high but decreasing with z, so the probability of grade school enrollment will drop with age.

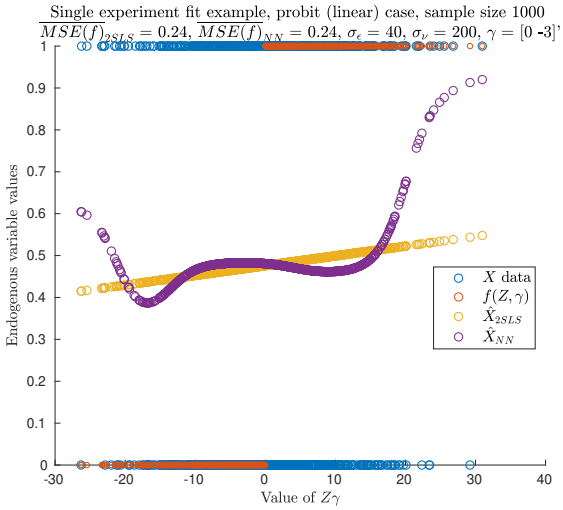
Other relationships may not be linear. Suppose our x variable is minivan ownership, x^* the corresponding perceived utility, and let z again be age. Young people likely have a low utility of minivan ownership. As they get older and start a family, a minivan becomes more useful. Once the children grow up, the utility may once again fall. In this example, $f(\cdot)$ may be closer to quadratic, but is certainly not linear.

We begin by considering a linear DGP for the x^*. Figure 2.9 demonstrates the performance of the linear and NN approaches in a high and low noise scenario. We graph the probability of observing 1 as estimated by both methods; these are the $\hat{X}$. We also plot the binary X data as observed, and what it would be in the noiseless case (labeled as $f(Z, \gamma)$). One can think of this as having perfect information regarding the value of school enrollment in the previous example, so that enrollment occurs until the age at which the value becomes negative, after which it ceases.

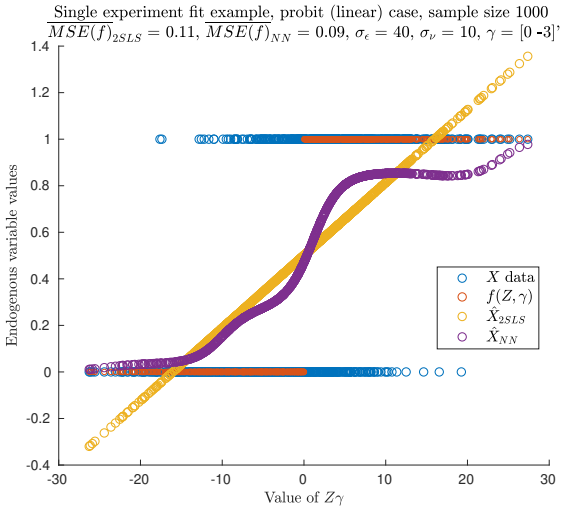
The results are similar to the linear case discussed in section 2.5.1. In the noisy case ($\sigma_\nu = 200$), neither approach performs very strongly and both are biased towards the OLS estimate, though for the first time $\hat{\beta}_{2SLS}$ has a slightly lower variance than $\hat{\beta}_{NN}$ (figure 2.9(c)). When the odds of observing a 1 are nearly 50-50 at every instrument value, both methods find it difficult to disentangle the true signal from the noise, resulting in widely varying estimates that frequently misidentify the direction of the OLS bias.

Both approaches perform better when less noise is present ($\sigma_\nu = 10$), rendering the instrument stronger. As before, the neural network manages to fit the first stage signal more closely. $\hat{\beta}_{NN}$ is slightly more biased than is $\hat{\beta}_{2SLS}$, but has a lower variance, which results in a lower mean squared error of the estimator. While we report results only for sample size 1000, the intuition regarding how performance varies with sample size developed in section 2.5.1 continues to apply.

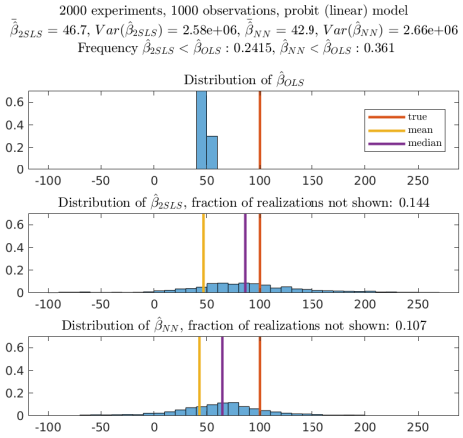
In figures 2.10, 2.11, and 2.12 we report the results for more complex first stage probit DGPs. Each case shown is the result of 2000 experiment iterations, with a sample size of 1000 in each case. As before, the linear approach is very poorly tailored to handle the



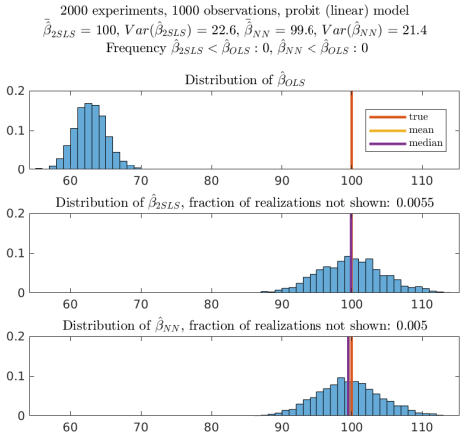
(a) Probit (linear) DGP, $\sigma_\nu = 200$



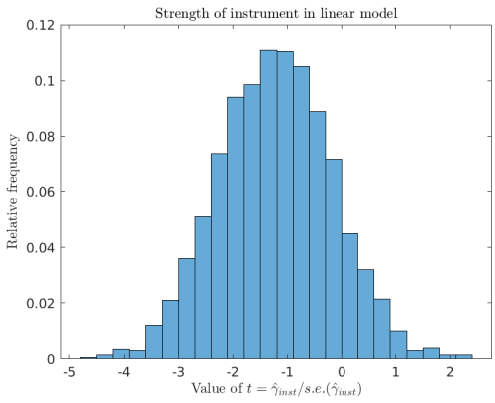
(b) Probit (linear) DGP, $\sigma_\nu = 10$



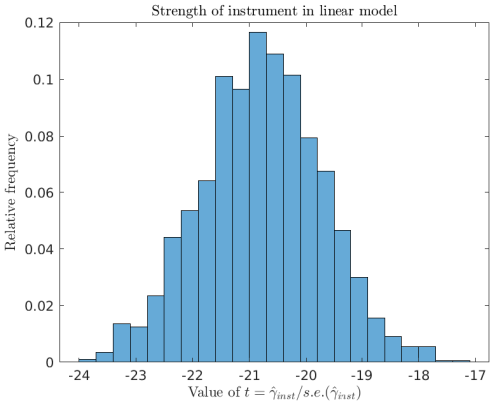
(c) Probit (linear) DGP, $\sigma_\nu = 200$



(d) Probit (linear) DGP, $\sigma_\nu = 10$

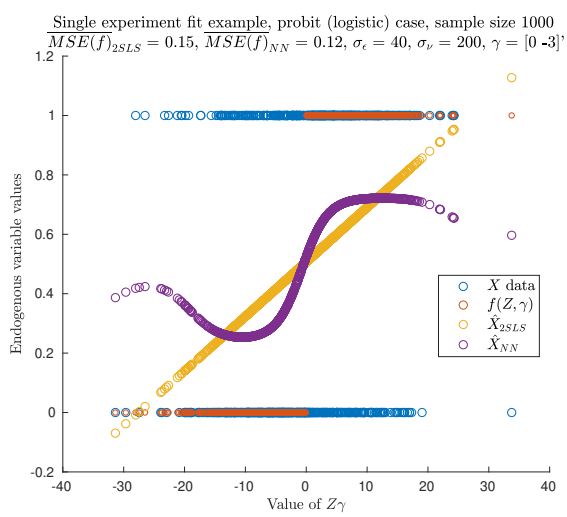


(e) Probit (linear) DGP, $\sigma_\nu = 200$

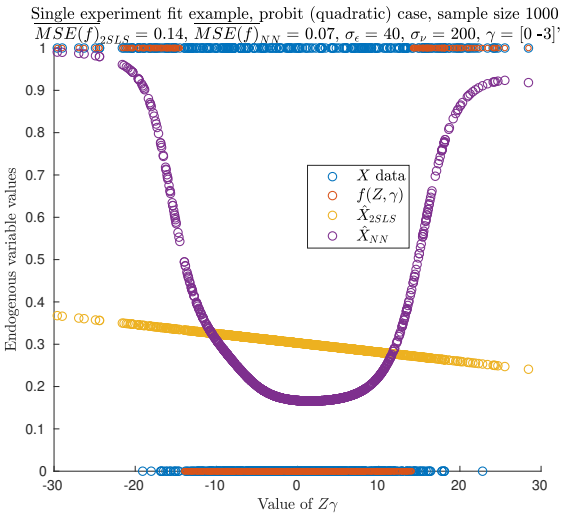


(f) Probit (linear) DGP, $\sigma_\nu = 10$

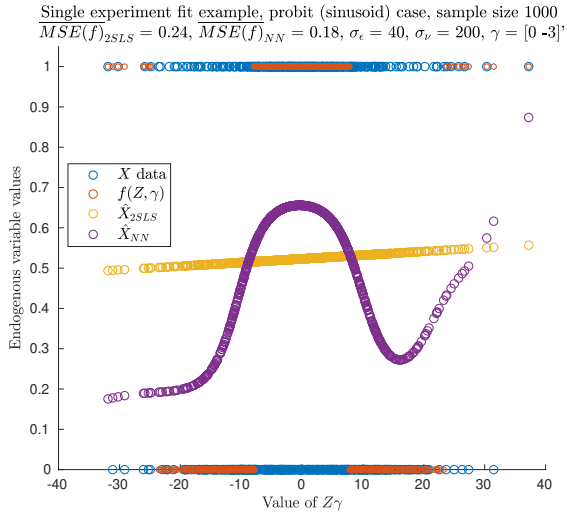
Figure 2.9: Dummy endogenous regressor with linear latent DGP (normally distributed ν). The figures in the left column, 2.9(a), 2.9(c), and 2.9(e), show results for $\sigma_\nu = 200$; the figures in the right column show results for $\sigma_\nu = 10$. In both cases, $x_i = 1$ is observed if $x_i^* = \gamma_0 + \gamma_1 z_i + \nu_i > 0$, and $x_i = 0$ otherwise. The red dots show the "true" x_i if ν_i were 0.



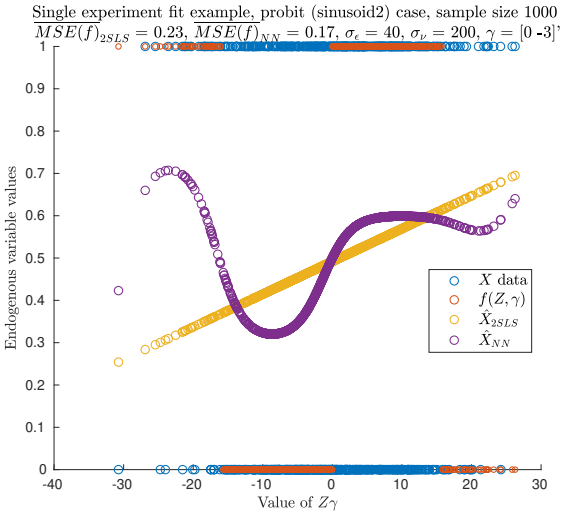
(a) Probit (logistic) DGP



(b) Probit (quadratic) DGP

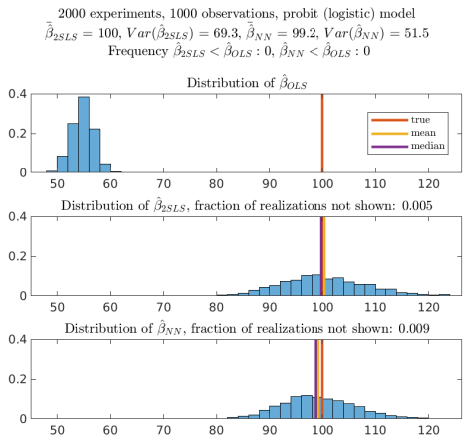


(c) Probit (sinusoid) DGP

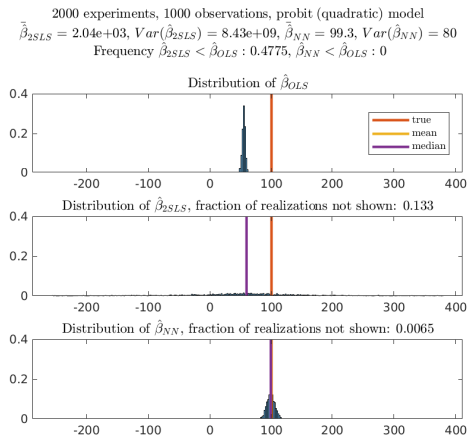


(d) Probit (sinusoid2) DGP

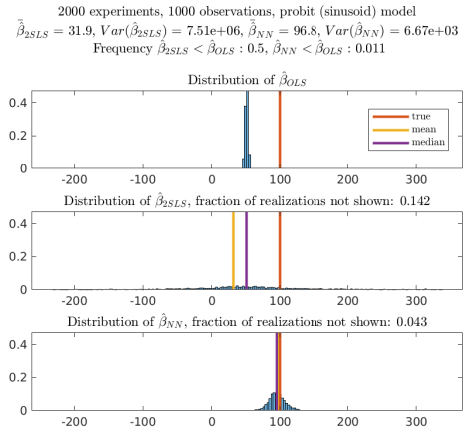
Figure 2.10: Dummy endogenous regressor with latent DGPs described by various functions. $x_i = 1$ is observed if $x_i^* = f(z_i, \gamma) + \nu_i > 0$; otherwise, $x_i = 0$ is observed. The figure shows example first stage fits from a single experiment iteration with different first stage DGPs. The red dots show the "true" x_i if all ν_i were 0.



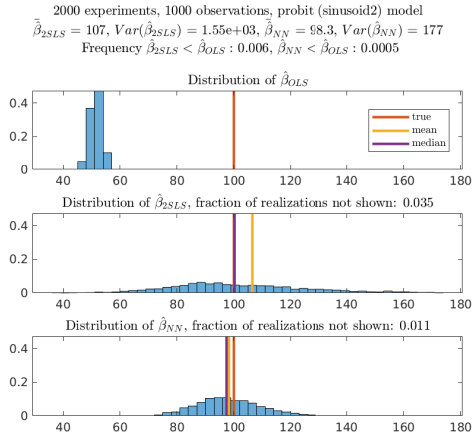
(a) Probit (logistic) DGP



(b) Probit (quadratic) DGP

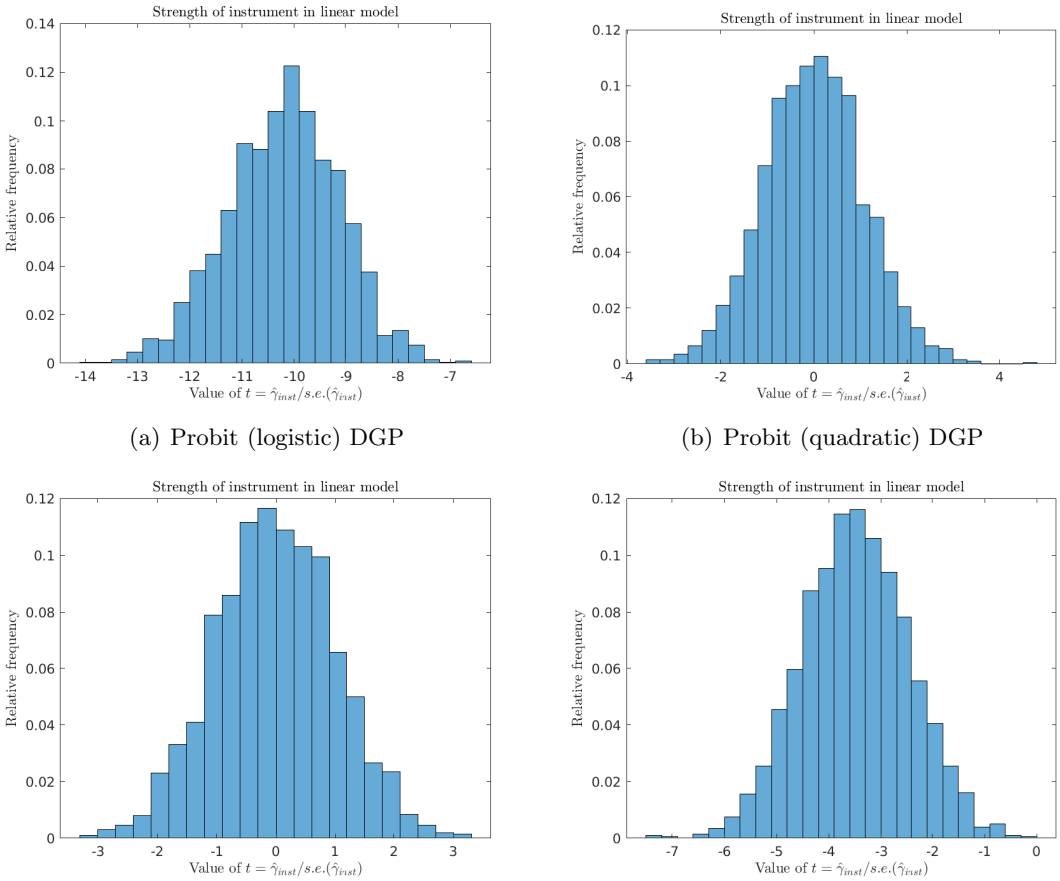


(c) Probit (sinusoid) DGP



(d) Probit (sinusoid2) DGP

Figure 2.11: Dummy endogenous regressor with latent DGPs described by various functions. $x_i = 1$ is observed if $x_i^* = f(z_i, \gamma) + \nu_i > 0$; otherwise, $x_i = 0$ is observed. The figure shows the values of $\hat{\beta}$ achieved using each estimation method.



(a) Probit (logistic) DGP

(b) Probit (quadratic) DGP

(c) Probit (sinusoid) DGP

(d) Probit (sinusoid2) DGP

Figure 2.12: Dummy endogenous regressor with latent DGPs described by various functions. $x_i = 1$ is observed if $x_i^* = f(z_i, \gamma) + \nu_i > 0$; otherwise, $x_i = 0$ is observed. The figure shows the t values when testing the strength of the instrument in the first stage of 2SLS.

quadratic and sinusoid (cosine) nonlinearities. (Certainly, the researcher could transform the variable prior to applying the linear method, but that requires a guess of the functional form, which it is unclear how to determine and is unnecessary when using the NN approach.) In both of those cases, the $\hat{\beta}_{2SLS}$ are much more biased and have a higher variance than do the $\hat{\beta}_{NN}$ (figures 2.11(b) and 2.11(c)). This is because when estimated linearly, the instrument appears weak (figures 2.12(b) and 2.12(c)). Because there is more than a single switch in the (de-noised) dummy variable, 0 to 1 and back, over the domain of the instrument, the linear method struggles – as would standard probit, for that matter, unless the instrument were correctly transformed before being used in the estimation.

In the remaining two cases, 2SLS manages to produce reasonable results, though it still performs best when there is only a single switch of the (de-noised) dummy variable from 0 to 1, as occurs in the logistic cdf DGP case. Here, as before, the $\hat{\beta}_{NN}$ is slightly more biased towards the OLS estimate, but this is more than made up for by a decrease in variance, leading to an overall lower mean squared error of the estimator.

When the $f(\cdot)$ is as specified for sinusoid2, 2SLS does manage to find that there is a relationship between z and x, but it is relatively weak (figure 2.12(d)). The $\hat{\beta}_{2SLS}$ estimates are thus both more biased, actually overestimating the true parameter value on average, and much more spread out than the $\hat{\beta}_{NN}$. They are also slightly more likely to produce a point estimate of β which is lower than that produced by OLS.

2.6 Discussion: extensions

Our results indicate that exploring the possible nonlinearity in the relationship between an instrument and the regressor for which it is being used can be worthwhile. In most cases, the gain is in the efficiency of the estimators, but as we have shown, unaccounted-for nonlinearity can also lead a linear method astray. A few more components are required, however, before the method can be applied in practice on non-simulated data.

2.6.1 Variance

In reporting our results, we rely on the empirical variance observed across samples, as we consider it the clearest and most accurate way to describe the comparison between methods. When applying these methods in practice, multiple samples are not available, so a formula is required to calculate the variance of the estimator. For standard 2SLS,

assuming homoskedasticity, it is the following:

$$\text{AVar}(\hat{\beta}_{2SLS}) = \sigma_\epsilon^2 (X'Z(Z'Z)^{-1}Z'X)^{-1} \tag{2.5}$$

The above result holds when using Z as an instrument. When we estimate $\hat{\beta}_{NN}$, we use $\hat{X}_{NN} = \hat{f}$ as an instrument. So long as $\hat{f}$ is a valid instrument, the same formula as above will hold. If Z and ϵ are independent, as they are in our simulations, then any function of the instrument should also be independent of ϵ, resulting in $\hat{f}$ being a valid instrument. Independence is a strong requirement, but if it is met, the derivation proceeds as usual:

$$\begin{aligned}
&\sqrt{n}(\hat{\beta}_{NN} - \beta) = \sqrt{n}(\{\hat{f}'X\}^{-1}\hat{f}'\epsilon) = \left(\frac{\hat{f}'X}{n}\right)^{-1}\frac{1}{\sqrt{n}}\hat{f}'\epsilon \\
&\text{plim}\left(\frac{\hat{f}'X}{n}\right) = \mathbb{Q}_{FX}{}^{*} \\
&\text{plim}\frac{1}{\sqrt{n}}\hat{f}'\epsilon = \text{plim}\sqrt{n}\cdot\frac{1}{n}\sum_i \hat{f}_i\epsilon_i \\
&\text{Var}[\hat{f}_i\epsilon_i] = \mathbb{E}[(\hat{f}_i\epsilon_i - \mathbb{E}(\hat{f}_i\epsilon_i))(\hat{f}_i\epsilon_i - \mathbb{E}(\hat{f}_i\epsilon_i))'] = \mathbb{E}[(\hat{f}_i\epsilon_i)(\hat{f}_i\epsilon_i)'] = \mathbb{E}[\hat{f}_i\epsilon_i\epsilon_i'\hat{f}_i'] = \\
&= \mathbb{E}[\hat{f}_i\ \mathbb{E}\{\epsilon_i^2|Z\}\ \hat{f}_i'] = \sigma_\epsilon^2\mathbb{E}[\hat{f}_i\hat{f}_i'] = \sigma_\epsilon^2\mathbb{Q}_i \qquad \Rightarrow \qquad \text{Var}\left[\frac{1}{n}\sum_i \hat{f}_i\epsilon_i\right] = \sigma_\epsilon^2\mathbb{Q}
\end{aligned} \tag{2.6}$$

So long as the data are well-behaved and no particular $\mathbb{Q}_i$ dominates, the last line will hold true. We can then apply the Lindeberg-Feller central limit theorem to say that:

$$\begin{aligned}
\left(\frac{1}{\sqrt{n}}\right)\hat{f}'\epsilon \xrightarrow{d} N[0, \sigma_\epsilon^2\mathbb{Q}] \quad \Rightarrow \\
\mathbb{Q}_{FX}^{-1}\left(\frac{1}{\sqrt{n}}\right)\hat{f}'\epsilon \xrightarrow{d} N[0, \sigma_\epsilon^2\ \mathbb{Q}_{FX}^{-1}\ \mathbb{Q}\ \mathbb{Q}_{XF}^{-1}]
\end{aligned} \tag{2.7}$$

Rearranging the expression for the variance, we can rewrite it as:

$$\text{AVar}(\hat{\beta}_{NN}) = \sigma_\epsilon^2 (X'\hat{f}\ (\hat{f}'\hat{f})^{-1}\hat{f}'X)^{-1} \tag{2.8}$$

It may be possible to relax the independence assumption; we leave this for future research. In the simulations, the variance values calculated in each sample are close to the variance of the distribution across samples.

[§]The usual conditions apply; Q_{FX} must have full rank. This will occur if each exogenous variable in the model acts as its own instrument, and each endogenous variable is predicted in such a way that it is not a linear combination of other variables.

2.6.2 Statistical significance

Although it is not a foolproof way of establishing the relevance of an instrument, statistical testing of the significance of the coefficients in a linear first stage can reassure the researcher that some relationship exists. In estimating a potentially nonlinear relationship using a neural network, we do not have an obvious way of testing whether the relationship is significant. Several possibilities exist to check for a statistical relationship, but we are not aware of a test of nonlinear significance analogous to the F test.

One approach is to calculate a "nonlinear correlation" between instrument and endogenous regressor, such as the maximal information coefficient, or MIC (Reshef et al., 2011; Speed, 2011). The MIC falls within the range [0,1]. 0 represents no relationship between two variables, and 1 corresponds to some kind of noise-free relationship, which does not have to be linear. This provides a useful statistic, but unfortunately does not hint at the distribution of this statistic in the population, so there are no standard errors to calculate which would suggest whether the point estimate is different from zero. It may be possible to develop the necessary theory, but a potential approach in the meanwhile is to perform a bootstrap estimation to approximate this distribution.

Another alternative is to test the parameters that the neural network training algorithm determines. Once the model is estimated, it is a particular nonlinear functional form. Knowing the shape of the relationship can allow the researcher to re-estimate it using methods that allow for a standard error to be calculated. Unfortunately, neural networks (as they are currently understood and implemented) do not produce statistics one can use to determine statistical significance directly.

Finally, it is the nonlinearity that is important to detect. A neural network approach is only one way in which to do it. For a simple model, then, it is possible to use the neural network estimation for exploratory analysis. This allows the researcher to initially remain fairly agnostic about the shape of the relationship between instrument and regressor, but later to develop a hypothesis. If one can visualize the relationship the neural network predicts, an obvious improvement on the standard linear approach may become apparent. For instance, including a quadratic term may clearly improve fit, or perhaps a periodic relationship will be detected. In that case, the researcher can transform the data based on the resulting hypothesis prior to running standard linear estimation, which would then estimate the imposed relationship and allow for straightforward significance testing.

2.6.3 More complicated models

There are further interesting extensions that can be explored, which we leave for future research. For instance, does the gain in efficiency persist in more complicated models with numerous endogenous variables and instruments? How is performance affected if errors are not symmetric but instead positively or negatively skewed? Another interesting consideration is whether we even need an instrument if we estimate the first stage nonlinearly.

Angrist and Pischke (2009) reference this possibility, but dismiss it as uninterpretable. Potentially, there can be an interpretation. It is not always necessary to find instruments outside of the model being estimated; for example lags are commonly used as instruments when estimating time series regressions. If we omit overtly including an outside instrument when estimating the first stage and instead predict the endogenous variable using some nonlinear combination of exogenous regressors, this is not so different from using a lag of a variable as an instrument – although of course there should be a reasonable underlying theory as to why a nonlinear function of some of the variables would shift the endogenous regressor. It may, however, be possible that the square of one of the variables could be used as an instrument for a different variable, if it is not already present in the regression.

2.7 Conclusion

In this paper, we present evidence of the impact a nonlinear underlying DGP in the first stage of 2SLS can have on the estimation of the parameters of interest in the second stage. While in some cases an undetected first stage relationship is simply a missed opportunity, and the researcher can proceed using a different instrument, in other cases the linearly-estimated relationship can appear statistically significant while incorrectly estimating the true DGP, resulting in misleading inference. Neural networks provide a way to explore potential nonlinearity while avoiding the constraint of imposing a specific set of functional forms – researchers can rely on their great flexibility and the property of universality to trace out a wide class of functions. Further research is needed to determine what aspects of the estimation method influence the performance of neural network based IV estimates of second stage parameters, but we believe it to be a promising approach.

References

Joshua D. Angrist and Jörn-Steffen Pischke. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press, 2009.

Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

Richard Blundell, Xiaohong Chen, and Dennis Kristensen. Semi-nonparametric IV estimation of shape-invariant Engel curves. *Econometrica*, 75(6):1613–1669, 2007.

Léon Bottou. Large-scale machine learning with stochastic gradient descent. *Proceedings of COMPSTAT'2010*, pages 177–186, 2010.

Xiaohong Chen and Demian Pouzo. Estimation of nonparametric conditional moment models with possibly nonsmooth generalized residuals. *Econometrica*, 80(1):277–321, 2012.

Steven Dieterle and Andy Snell. It's hip to be square: Using quadratic first stages to investigate instrument validity and heterogeneous effects. *Unpublished manuscript, Edinburgh: University of Edinburgh. Retrieved from: http://homepages.econ.ed.ac.uk/ sdieterl/researchpapers/DieterleSnellIV.pdf*, 2014.

Ethan Dyer and Guy Gur-Ari. Asymptotics of wide networks from feynman diagrams. *Proceedings of the 36th International Conference on Machine Learning, ICML*, 97, 2019.

Jason Hartford, Greg Lewis, Kevin Leyton-Brown, and Matt Taddy. Deep IV: A flexible approach for counterfactual prediction. *Proceedings of the 34th International Conference on Machine Learning, PMLR*, 70:1414–1423, 2017.

Jerry A. Hausman. An instrumental variable approach to full information estimators for linear and certain nonlinear econometric models. *Econometrica*, 43(4):727–738, 1975.

Joel L. Horowitz. Applied nonparametric instrumental variables estimation. *Econometrica*, 79(2):347–394, 2011.

Hamed Masnadi-shirazi and Nuno Vasconcelos. On the design of loss functions for classification: theory, robustness to outliers, and SavageBoost. *Advances in Neural Information Processing Systems*, 21, 2008.

Sendhil Mullainathan and Jann Spiess. Machine learning: An applied econometric approach. *Journal of Economic Perspectives*, 31(2):87–106, 2017.

Whitney K. Newey. Semiparametric efficiency bounds. *Journal of Applied Econometrics*, 5:99–135, 1990.

Whitney K. Newey and James L. Powell. Instrumental variable estimation of nonparametric models. *Econometrica*, 71(5):1565–1578, 2003.

Whitney K. Newey, James L. Powell, and Francis Vella. Nonparametric estimation of triangular simultaneous equations models. *Econometrica*, 67(3):565–603, 1999.

David N. Reshef, Yakir A. Reshef, Hilary K. Finucane, Sharon R. Grossman, Gilean McVean, and Peter J. Turn. Detecting novel associations in large data sets. *Science*, 334(6062):1518–1524, 2011.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951.

Terry Speed. A correlation for the 21st century. *Science*, 334(6062):1502–1503, 2011.

Alwyn Young. Consistency without inference: Instrumental variables in practical application. *Unpublished manuscript, London: London School of Economics and Political Science. Retrieved from: http://personal.lse.ac.uk/YoungA/ConsistencyWithoutInference.pdf*, 2019.

Chapter 3

Suboptimal Decision-Making on Stochastic Lotteries in Indonesia

3.1 Introduction

What is preferable, £1500 per month with certainty, or a gamble to earn either £1500 or £3000 per month with equal probability? A rational economic actor would find this question trivially simple; yet 42% of respondents to the 2007 Indonesia Family Life Survey may disagree. How could so many respondents answer such a fundamental question to good economic decision-making incorrectly?

Recent studies have observed numerous behavioral anomalies in which economic theory based on standard rationality assumptions fails to explain the behavior of individuals (e.g. Camerer and Loewenstein, 2004; Kahneman and Tversky, 1979; Shogren and Taylor, 2008; McFadden, 1999). Rabin and Thaler (2001) argue that people are closer to being myopic loss-averters than expected utility maximizers because they seem to assess risk differently for small and for large gambles. Agents do not display approximate risk-neutrality at the level of small gambles, which we would observe if standard expected utility theory characterized their decision-making. Gneezy et al. (2006) observe what they call the "uncertainty effect". In a series of experiments, they show that groups of individuals on average value a gamble with two possible outcomes less than they value each individual outcome, violating internality. Ambiguity aversion is another behavioral anomaly first noted by Keynes (1921) and later elaborated upon by Ellsberg (1961) and others. It describes individuals' preferences towards clearly-defined known risks (known probabilities of various outcomes) rather than unknown risks, even in situations where, according to stan-

dard economic theory, it does not matter. Kőszegi and Rabin (2006) develop a theory to explain unexpected behaviors such as the endowment effect, which is sometimes observed in individuals, explaining it as a matter of the expectations people hold. This paper seeks to extend the literature by applying a modified version of Kőszegi and Rabin's theory to explain why individuals may turn down an option which on the surface may seem like an obvious improvement.

Respondents in Indonesia were presented with a choice between a sure option and a *strictly favorable* gamble (a purely hypothetical choice with no real money involved), and a large number of them chose the sure option. Specifically, in the 2007 Indonesia Family Life Survey (IFLS) (Strauss et al., 2009), Book 3a, Section SI, respondents are asked:

> **Question 1** *Suppose you are offered two ways to earn some money. With option 1, you are guaranteed Rp. 800 thousand per month. With option 2 you have an equal chance of either the same income, Rp. 800 thousand per month, or, if you are lucky, Rp. 1.6 million per month, which is more. Which option will you choose?*

If the respondent chooses the certain option, he or she is prompted to reconsider, to make sure that option 1 is truly his or her preference:

> *Are you sure? In option 2 you will get at least Rp. 800 thousand per month and you may get Rp. 1.6 million per month. In option 1, you will always get Rp. 800 thousand per month.*

To anyone familiar with economic theory, option 2 clearly dominates option 1. Yet 45% of respondents choose option 1, and less than 8% of those individuals change their mind when asked if they are sure, which amounts to over 40% of the full sample choosing and sticking with option 1. One could say respondents simply don't understand the question and guess at the answer. Perhaps if real money were involved and respondents previously observed transactions occurring, they would choose differently. However, there is little reason for them to lie about their preferences — if so, why would 60% of respondents still choose option 2? — and it seems unlikely that they are choosing randomly, as answers to other similar questions do not display such an erratic pattern. Given the large number of people choosing each option, as well as the observation that a significant minority of even the most highly-educated people in the sample choose option 1, it seems there is some intentionality in their answers.

The explanation I develop in this paper is based on an application of Kőszegi and Rabin's (2006) reference-dependent utility. I hypothesize that individuals choose as they do due to a combination of aversion to losses and anchoring on the tempting higher outcome. If the higher outcome is not realized, it is interpreted as a loss, and the individual is disappointed. In order to avoid such disappointment, some individuals choose to avoid the lottery, preferring the certain option with its lack of surprises.

In section 2, I discuss possible alternative explanations and other influences which may impact the choice observed in Question 1 alongside the modeled explanation. Section 3 presents the model. Section 4 estimates the effect of individual characteristics or life circumstances which may influence the person's decision-making. Section 5 discusses the potential implications of the observed behavior on the lives of the individuals choosing the dominated option. Section 6 presents possible avenues for future research and discusses policy implications. Section 7 concludes.

3.2 Possible Explanations

There may be numerous reasons why the respondents choose as they do. Self-aware respondents may be wary of the higher-income money-making opportunity if they think they would be unable to put the extra money to productive use, and instead spend it on alcohol, sugar, cigarettes, or similar "temptation goods". While we cannot decipher which individuals may be sophisticated hyperbolic discounters, we do observe the consumption of these goods at the household level, which can be interpreted as the degree to which members of the household give in to temptation. I construct a total weekly spend on sugar, soft drinks, alcohol, betel nut, cigarettes, and pre-prepared food. Controlling for an individual's income, a higher (household) consumption of temptation goods is actually associated with a slightly lower probability of choosing the dominated option[1]. The propensity to choose the dominated option does not seem to result from any attempts at self-control.

Respondents who live in unsafe circumstances may feel that having relatively too much money could draw attention to them, perhaps putting them in danger. The median monthly income earned among those reporting an income, 16,745 respondents, is Rp. 500,000 per month. Perhaps Rp. 800,000 per month is already quite generous in com-

[1]The effect on the choice is nearly negligible, less than 1 percentage point, but it is statistically significant with t=-2.49. Including only those households with one respondent leaves the estimate small and negative, but no longer statistically significant.

parison. I check if vulnerability may potentially explain the choice to turn down "excess" money by checking how the likelihood of this choice varies with subjectively-reported feelings of safety in the community. I find that there is no statistically significant difference between those respondents who feel safe (the majority) and those who do not.

Respondents may misunderstand or reinterpret the question. Since they are asked about "ways to earn some money", they may well consider the kinds of jobs that pay the salaries stated and project the characteristics of those jobs onto the question at hand. Alternatively, while the question states that there is an "equal chance" of either outcome, it is possible to interpret that statement as there being a probability p of each stated outcome and some probability $1 - 2p$ of receiving neither.

To account for the possibility that respondents are not paying attention or are not seriously responding to the survey, I check if those whom the interviewer judged to be inattentive are more prone to choosing the dominated option. That is indeed the case, but does little to explain how frequently the dominated option is chosen. Among those taking the survey seriously, which is most participants, 41.6% choose "incorrectly". Among the rest, the fraction is 43.4%. The difference is only statistically significant at the 10% level. Accounting for children or other adults being present during the interview yields similar results – greater potential for distraction corresponds to a statistically significant higher likelihood of choosing the dominated option, but the difference is slight.

A lack of education may lead individuals to reason in nonstandard or illogical ways. Having attended junior or senior high school is associated with a highly statistically significant 9 percentage point lower probability of choosing the dominated option relative to someone who is less educated. If the respondent has attended tertiary education, the effect is an even higher 19 percentage points. However, even among those who have attended tertiary education, 28.4% still choose the dominated option. It is possible that a lack of education contributes to misunderstanding the question and selecting the dominated option, but a lack of understanding does not seem to be the only channel which impacts the probability of choosing "incorrectly".

It is likely that framing also plays a role. In the question I describe, the extra money is framed as a gain one receives if one is "lucky", and 42% of respondents turn down the extra gain. The survey contains a similar later question framed as a loss:

Question 2 *Suppose you are offered two ways to earn some money. With option 1,*

you are guaranteed an income of Rp. 4 million per month. With option 2 you have an equal chance of either the same income, Rp. 4 million per month, or, if you are unlucky, Rp. 2 million per month, which is less. Which option will you choose?

Similarly to the previous question, respondents who choose the now-undesirable lottery in this scenario are prompted to reconsider, and many fewer respondents make a "mistake". Of the 28,904 respondents who answered the question, only 8% choose the second option and stick with it when asked a second time. Here, the Rp. 2 million per month outcome is presented as an unnecessary loss. While the difference in numbers used in the questions makes direct comparison not quite straightforward, it seems reasonable that the want to avoid a loss — the decrease in utility from which is accepted to be more than the increase in utility from a comparable gain — would drive more people towards making the rational choice according to economic theory than a want to pursue a gain.

3.3 Model

To explain the puzzle, I apply a version of Kőszegi and Rabin's (2006) model. They model utility as being reference-dependent, defining the utility function as: $u(c|r) = m(c) + \mu[m(c) - m(r)]$, where $m(c)$ is a usual concave utility function and $\mu[m(c) - m(r)]$ is an adjustment to the utility derived from consumption depending on how far it is from expectations. $\mu(\cdot)$ is a function with properties described in Kahneman and Tversky's prospect theory. I take the simplest form of the $\mu(\cdot)$ function:

$$\mu(x) = \begin{cases} x & x \geq 0 \\ \beta x & x < 0 \end{cases}, \text{ where } \beta > 1 \tag{3.1}$$

In the lottery example presented in Question 1, there are two possible outcomes, x and $2x$, where x corresponds to Rp. 800 thousand, and $2x$ corresponds to Rp. 1.6 million. Their alternative is a certain x. The reference point is modelled stochastically and can be different for each person. Individuals can anchor on x with probability p and they can anchor on $2x$ with probability $1 - p$. The individual's characteristics or past experience shape their value of p. An individual is thus fully defined by their $m(\cdot)$ function, p, and β.

Option 1, as presented in the question, guarantees a utility of $m(x)$. There are no other outcomes for which one might hope, so the reference point is the same certain outcome.

If one's reference point is x, the expected utility of Option 2 becomes:

$$\mathbb{E}(u_2|r = x) = .5m(x) + .5[m(2x) + \{m(2x) - m(x)\}] \tag{3.2}$$

If one's reference point is $2x$, the expected utility is the following:

$$\mathbb{E}(u_2|r = 2x) = .5[m(x) + \beta\{m(x) - m(2x)\}] + .5m(2x) \tag{3.3}$$

The overall expected utility of choosing Option 2 is thus:

$$\begin{aligned} \mathbb{E}(u_2) &= p \cdot \mathbb{E}(u_2|r = x) + (1 - p) \cdot \mathbb{E}(u_2|r = 2x) = \\ &= .5 \cdot \{(m(x) + m(2x)) + (m(2x) - m(x)) \cdot \underbrace{[p - (1 - p)\beta]}_{\theta}\} \end{aligned} \tag{3.4}$$

It makes sense to choose the certain option over the lottery if its expected utility is higher, in other words when $m(x) = \mathbb{E}(u_1) > \mathbb{E}(u_2)$. Comparing the two expressions and simplifying shows that this occurs when $\theta < -1$. This is equivalent to $p < \frac{\beta-1}{\beta+1}$. Given the constraints of the $\mu(\cdot)$ function, we know that β must be greater than 1; the case where it equals 1 corresponds to no loss aversion. As β increases and the individual becomes more and more loss averse, the right-hand side of the expression also increases, making it more likely that an individual's p falls below the threshold. If they are to choose the lottery, the increasingly loss averse individual must be more and more likely to anchor on x as their reference point, as opposed to $2x$, to avoid the pain of disappointment. If an individual is highly loss averse, putting even a small weight on the high outcome may be enough for them to prefer the certain option. Hoping to get $2x$ and then receiving only x is too disappointing.

We can perform the same analysis on Question 2. Here, the certain option is now $2x$ (Rp. 4 million), and the lottery is over x (Rp. 2 million) and $2x$ (Rp. 4 million). The certain option now dominates the lottery. The condition which ensures the agent will choose the lottery in this case is $\theta > 1$, which corresponds to $\beta < -1$. This is impossible, as β must be greater than 1. The model thus predicts that no one should choose the lottery in Question 2 due to their loss aversion. Indeed, nearly all respondents answer that question as we would expect. Those that choose the lottery must be doing so for some other reason, unless perhaps they are risk loving and find more pleasure in gains than pain in losses.

3.4 Estimation

We now have a framework through which to interpret the choice an individual is making. Some characteristics or circumstances may directly increase the probability of making an unintentional mistake, such as a lack of education, and some act through the channel of deliberate decision-making by a loss-averse agent. For example, one's past experience may leave one strongly averse to losses (high β) or may affect the option on which one anchors (p).

I now estimate how the likelihood of choosing the stochastically dominated option in Question 1 varies across people as their circumstances or their traits change, and discuss how these impacts may be interpreted in the disappointment aversion framework. The data unfortunately do not provide a measure of the respondents' loss aversion, nor is there any direct indication of how they may likely be forming their reference point. I use a simple regression to document the effects for easy interpretability[2], and I allow for household-level clusters in the error term.

Table 3.1 shows the results. We see that as before, whether the respondent is focused on the interview or not does not seem to play much of a role in determining their choice, so selecting the dominated option does not appear to be due to an inattentive mistake.

One may suppose that feeling a lack of safety may induce a greater degree of loss aversion, as those who feel vulnerable or precarious may suffer more when they incur a loss. Those who consider their community to be safe are indeed less likely to choose the dominated option, but this is only borderline statistically significant.

There is some evidence that we may become more loss averse as we age (Kurnianingsih et al., 2015), and we do see that the older the respondent, the more likely they are to choose the dominated option. Respondents in the sample range from 15 to 100 years old, and each decade increase in age is associated with a 1.7 percentage point increase in the probability of selecting the dominated option, even after controlling for education. This is a steady progression across the entire age support.

It appears women choose the dominated option more frequently than do men, even after controlling for feelings of safety in the community, level of education, and work activities. It is possible there may be cultural reasons why women are more likely to prefer the

[2]The resulting subgroups of the population tend to have 20%-60% of the group choosing the stochastically dominated option. Since we observe middle levels of the probability of choosing this option, a linear approximation should give reasonable estimates of the marginal effects without predicting probabilities outside of the [0,1] range.

Factors affecting decision-making		
	Coefficient	Std. Err.
Respondent attentive	-.0180983	.0156722
Community feels safe	-.0351520	.0217222
Age	.0017077***	.0003504
Female	.07482700***	.0082661
Attended Jr/Sr HS	-.0672507***	.0094423
Attended tertiary	-.1399806***	.0133777
High income	-.0625559***	.0091904
Primary daily activity:		
Not working	-.0159632	.0137188
In school	-.0805185*	.0431520
Self-emp. (no paid emp.)	-.0202286***	.0088575
Self-emp. ($\geq$ 1 paid emp.)	-.0745017***	.0254097
Constant	.4329111***	.0293526
n	15,594	

Table 3.1: Likelihood of choosing stochastically dominated option presented in Question 1 across various sub-populations. The education dummies correspond to highest level of schooling attended. The high income dummy is 1 for individuals whose annual income is Rp. 800,000 or higher and 0 otherwise. The omitted category of primary daily activity is salaried employment. Errors are clustered at the household level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

smaller, certain amount of money. Alternatively, it may be because women are more loss averse or more likely to anchor on the higher value, so the potential disappointment from the gamble makes the certain option more preferable. It has not yet been established whether men or women are more averse to losses (e.g. Bouchouich et al., 2019).

It was mentioned earlier that higher education is associated with a statistically significantly much lower probability of selecting the dominated option, and this remains the case after controlling for other factors, although the estimates decrease slightly. It seems likely that education could have a direct effect on how an individual considers questions such as Question 1, as it may be better understood and more likely to be regarded as an academic exercise with a correct and incorrect answer. A higher level of education may also induce the individual to consider choices more logically and less emotionally. If so, education could also potentially lower a person's level of loss aversion.

A higher annual income is associated with better decision-making. In particular, those who earn above Rp. 800,000 per month are 6.3 percentage points less likely to choose the dominated option, and this difference is statistically significant. Having a lower income may lead one to simultaneously feel greater distress at losses and to also anchor on the tempting high amount. This combination would then lead those with a lower income to avoid the gamble in order to avoid disappointment. It may also be the case that those

with a higher annual income can approach the gamble more dispassionately and form a reference point with $p = .5$ as opposed to focusing on the higher value.

Finally, I look at how people respond differently based on their main daily activity, taking salaried employment as the base category. Those who are not employed are not very different from those who are salaried in terms of their survey answer. Those who currently attend school are 8 percentage points more likely to answer "correctly", although that effect is only significant at the 10% level. Those who are self-employed without any employees or with a family member providing unpaid labor to their business are 2 percentage points less likely to choose the stochastically dominated option, significant at the 5% level. Those whose business is more established, as indicated by having at least one paid employee, are 7.4 percentage points less likely to choose irrationally than are the salaried, statistically significant at the 1% level. Perhaps running a business provides more practice in decision-making, showing that losses may be temporary and not so painful. In forming the reference point, a business owner may be more rational and consider both outcomes equally, as opposed to anchoring on one or the other. Alternatively, it may be that those who are more rational decision-makers choose to start a business or are able to expand the business enough to hire employees.

3.5 Implications

Data from the IFLS indicate that a large fraction of the Indonesian population may avoid risk at all costs – even when there is no risk discernible to a rational decision-maker. This may partly be due to misunderstanding, but a lack of understanding does not appear sufficient to fully explain the prevalence of this behavior. The choice to avoid gambles may have evolved to be a kind of rule-of-thumb for some individuals. Perhaps through the course of their lives they have learned that getting one's hopes up only sets one up for disappointment. A certain option not only guarantees the outcome, it also guarantees a lack of surprises and a lack of painfully unmet expectations. If one harbors the slightest hope of getting $2x$, the gamble will disappoint if one does not get it – and the joy of receiving the larger amount may be much less than the pain of not receiving it.

Particularly worrying is that those who are less educated or poorer appear to be more prone to such total risk-avoidance. If growing up poor makes losses more painful (perhaps because one is so close to having nothing), then the lesson learned each time expectations are not met is that one should not have hoped for the good outcome. If one does not hope

for the good outcome, then there is little point in trying for this outcome – may as well choose the certain x, or may as well try to maintain the status quo. Loss aversion may thus lead to up a poverty trap.

While responses to a single survey question may hold limited information, the observed choices could hint as to why we see people in developing countries rejecting or being hesitant to accept new measures which will "obviously" improve their lives. Farming technologies tend to spread slowly (Conley and Udry, 2010; Foster and Rosenzweig, 1995), water bleaching or filtration devices see limited take-up and use (Kremer et al., 2008), antimalarial bed nets are not used as much as they perhaps should be (Dupas, 2014). It is possible that descriptions of the benefits are poorly understood (similar to saying "equal chance" in the first question, the interpretation of which can be ambiguous). But if the explanations are sufficient, then the more the individual believes in the benefits and anchors on them, the greater looms the potential for disappointment. The individual may still be strictly better off choosing the new way according to standard economic theory, but if we account for the disappointment, she may in fact feel she would be worse off. This is in line with what prior research has suggested (Yesuf and Bluffstone, 2009; Liu, 2013).

What the IFLS survey shows is that these preferences may lead people to turn down what is economically an unambiguously better option. Not only may people under-invest in risky opportunities as a result, they may under-invest even in what we would commonly consider to be *safe* options.

The IFLS survey contains a number of measurements of household assets, including the total rupiah value of "savings / certificate of deposit / stocks" owned by the household. We would expect that being loss averse to the extent that one would choose a stochastically dominated certain option over a lottery would lead one to under-invest in these assets, especially something as uncertain as stocks. As Table 3.2 shows, this is indeed the case.

The relevant coefficient is labeled "Option 1". This is a dummy variable indicating the respondent chose the dominated certain option over the lottery in Question 1. The outcome is a binary indicator for having any savings or investments in the household. As one might expect, those who are wealthier, older, and more educated are more likely to have some savings or invest in the stock market. It appears that women are also more likely to record a positive response, even though they more frequently choose the dominated option. Perhaps this is because the available measurement combines both savings and investment. While loss aversion may lead women to under-invest, it may not have that effect on their savings.

Impact of loss aversion on savings/investment

	Coefficient	Std. Err.
Option 1	-.0241527***	.0084723
Log annual income	.0514719***	.0035041
Age	.0010426***	.0003291
Female	.0788371***	.0069032
Attended Jr/Sr HS	.1066841***	.0098865
Attended tertiary	.3199161***	.0167967
Community feels safe	-.001899	.0251086
Constant	-.6437643***	.0585025
n	12,721	

Table 3.2: The outcome variable is a binary indicator for having any savings/investments in the household. Option 1 corresponds to choosing the stochastically dominated certain option over the lottery in Question 1. The education dummies correspond to highest level of schooling attended. Errors are clustered at the household level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Controlling for the above characteristics, as well as community safety, those who choose the dominated option are 2.4 percentage points less likely to have any savings or investments. This may be due to loss aversion constraining their investment, or it may be due to generally poor decision-making skills. The fraction of people in the sample who have any savings or investments is 26.8%. A decrease of 2.4 percentage points thus corresponds to a 9% decrease.

I also look into the amount reported under "savings / certificate of deposit / stocks". Conditional on reporting that the household owns any of these, those who choose the dominated option report a smaller amount. The mean amount saved or invested among those who choose "correctly" is Rp. 7,080,557, and among those who choose "incorrectly" it is Rp. 5,697,038. The difference between the two groups is significant at the 5% level. It thus appears that not only is choosing the dominated option associated with a lower likelihood of having any savings or investments, but it is also associated with a lower amount saved or invested if one does own these assets.

While establishing a causal relationship given the data constraints is not possible, I now look at the possible impact of nonstandard/irrational decision-making on people's employment choices. In particular, I build on the insights generated in the first chapter of this thesis and look at selection into entrepreneurship. While it is not possible to determine whether businesses are incorporated, I separate the businesses into "self-employed", which corresponds to a one-person business, potentially helped by an unpaid family member, and "entrepreneurs", who have at least one paid employee.

The IFLS questions contain four further decision-making questions which can be used to

	Entrepreneurship	Self-employment
Risk tolerance	.0038097***	.0141824***
	(.0009614)	(.0025257)
Option 1	.0019494	.0237007***
	(.0029169)	(.0074884)
Attended Jr/Sr HS	.0083919***	-.1582432***
	(.0020963)	(.0054067)
Attended tertiary	.0205941***	-.2413278***
	(.0031721)	(.0084824)
Female	-.0093222***	-.1535218***
	(.0019950)	(.0050377)
Constant	.009454***	.3954978***
	(.002938)	(.0078124)
n	18,680	26,792

Table 3.3: The outcome variable in each regression is a dummy for the type of employment. Option 1 corresponds to choosing the stochastically dominated certain option over the lottery in Question 1. The education dummies correspond to highest level of schooling attended. Risk tolerance is based on five bins taking values 0 to 4, which the latter corresponding to higher risk tolerance. Standard errors are reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

rank respondents from least to most risk tolerant. I thus assign values 1 through 4 to respondents, correspondingly. Necessarily, those who choose the stochastically dominated option in Question 1 must be assigned a risk tolerance value of 0; they are not asked the other four questions.

For each type of business ownership, I regress the dummy for selecting this occupation on education (as an imperfect proxy for ability), the risk tolerance measure, a dummy for being female, and a binary indicator of having chosen the dominated option (labeled Option 1). This allows for a further impact aside from through lowered risk tolerance, such as if the group of people who choose the certain option are unusually loss averse or more likely to anchor on high outcomes.

As expected, having a higher risk tolerance or being more educated is positively and statistically significantly associated with entrepreneurship. The opposite is true for being female. Indeed, there are only 93 female entrepreneurs to 250 male entrepreneurs. The coefficient on Option 1 is not statistically significant. It does not appear that the factors which lead an individual to choose the dominated option have any additional impact on the likelihood of being an entrepreneur beyond what is controlled for by the other variables.

When it comes to self-employment, however, we observe a qualitatively different set of results. While the coefficients on risk tolerance and education remain statistically significant, the coefficients on education are now negative. Those who are educated are much less likely

to be running a small, one-person business. This is similar to what Levine and Rubinstein (2017) find in their work comparing incorporated and unincorporated entrepreneurs in the US. In addition, the coefficient on Option 1 is now positive and statistically significant. Those who choose the dominated option are more likely to be running a one-person business, even though their risk tolerance is low. This makes sense, as such businesses are often the result of people having no other options to sustain themselves. It is possible a high degree of loss aversion or otherwise impaired decision-making makes it difficult to find a would-be-preferred salaried position. Alternatively, it may be that living in precarious circumstances impacts both an individual's preferences and their economic options. As in the case of entrepreneurship, being a woman is associated with a lower probability of running one's own business.

3.6 Future research

Two factors influence an individual's choice, according to the model. One is the individual's reference point and the other is their sensitivity to disappointment. In practice, there are likely other additional channels which shape the person's decision-making. To fully rule out misinterpretation, it would be helpful to gather more data while very clearly conveying the set-up in the question to individuals in a concrete way that is difficult to misinterpret. The question has already been thoughtfully designed, but perhaps respondents still don't fully grasp that, for instance, there is no outcome in which the lottery in Question 1 results in nothing. For example, Keren and Willemsen (2009) show that while they can replicate the uncertainty effect as first noted by Gneezy et al. (2006), in their sample it primarily comes from individuals misunderstanding the instructions. Rephrasing the question to make the statement on probability very clear by describing it using a coin flip resulted in participants' responses no longer demonstrating the uncertainty effect.

Once the lack of understanding channel has been eliminated, one could explore the extent to which loss aversion or reference point formation plays a role. According to the model prediction, it should make no difference whether the tempting higher option in the question is $2x$ or $3x$, if the respondent continues to have the same p and β, propensity to anchor on the lower option and the degree of loss aversion, respectively. Including this change is unlikely to influence the respondent's aversion to losses, but it may nudge respondents towards a greater propensity towards anchoring on the higher outcome. For any given level of loss aversion, individuals would then be more likely to turn down the gamble.

To assess the extent to which loss aversion leads people to avoid potential disappointment, one could measure the loss aversion of the respondents separately and then see if those who report a greater degree of loss aversion are also more likely to choose the certain option over the gamble in Question 1. Alternatively, one could rephrase the question to try to make use of individuals' loss aversion to nudge them towards the economically beneficial answer and then see how responses change. We can do this by placing the outcomes in the realm of potential losses, such as the following:

> *Suppose you are offered a way to earn some money which guarantees you Rp. 1.6 million per month. However, it requires you to choose one of the following two options. With option 1, you are will certainly lose (have to pay back) Rp. 800 thousand of that income. With option 2, you have an equal chance of either losing Rp. 800 thousand, or, if you are lucky, not losing any of the income. Which option do you choose?*

By taking a different perspective on the same outcomes, it may be possible to encourage more rational decision-making in the respondents.

If disappointment aversion truly does drive individuals to reject risky gambles, even when there is no risk involved in the traditional sense, this leads to some specific policy implications. As an example let's take the situation observed by Liu (2013), where farmers had the opportunity to adopt a genetically modified variant of cotton, which was advertised as having an increased yield and requiring a lower amount of pesticide (lower cost), all while being no riskier than the standard variety in terms of yield risk. Even so, some farmers waited up to a decade to switch to the new technology. This example can be mapped onto the set-up in Question 1 – either you choose to stick with the status quo, which is x, or you try switching to the new variety. True, the new variety may give you better profits, but what if it doesn't work out and you still get the same as before? That would be disappointing. As predicted by the model, more loss averse farmers adopt the new technology later than those less loss averse.

If we wanted to encourage adoption of this new technology, a simple change in the messaging may help. Instead of stating what the farmers stand to potentially gain, the phrasing could emphasize what they are losing out on by not making the switch. Presenting the switch as the default option and the lower yield as losses may encourage greater, faster adoption, as loss aversion may nudge farmers towards rather than away from adoption.

Another potential approach could be to guarantee some payment if a farmer tries the new

variety. In terms of the model, this would make the choice x vs equal chances of $x+y$ and $2x+y$, where y is the subsidy offered. The need for such a subsidy may seem preposterous if one is dealing with rational decision-makers, but perhaps it is necessary to help some individuals overcome their aversion to disappointment.

3.7 Conclusion

The IFLS survey conducted in Indonesia in 2007 presents us with some unexpected results. When nearly half of the sample selects a stochastically dominated option, it is difficult to write it off as simply noise in the data, some random mistakes. Yet even if these are mistakes, the magnitude observed makes it important to understand why people choose as they do and to account for this in policies which rely on individuals' ability to make "obviously" better, rational choices for themselves. In light of other behavioral economics research that has documented psychological considerations that markedly affect decision-making, from the way scarcity can affect perceived value (Shah et al., 2015) to preferences that indicate hyperbolic discounting (Dasgupta and Maskin, 2005), perhaps the choices observed in the IFLS survey are not so surprising.

I believe the responses recorded are thoughtful and intentional for the majority of respondents. It is likely that some respondents do not fully understand the question posed, but this does not seem to be enough to explain the scale of the "mistake" observed – nearly half the sample, including some of the most highly-educated individuals. This paper proposes that individuals choose the dominated option based on a high degree of loss aversion in combination with anchoring on the desired high outcome. It may seem prudent or sensible for a person to try to avoid disappointment, but it might lead them to turn down opportunities which any rational decision-maker would find obviously worth accepting. Of course, there may be other possible explanations. Further research is needed to more fully understand the process by which individuals make the kinds of choices exemplified in the IFLS question, and to take this decision-making process into consideration when designing policy interventions.

References

Ranoua Bouchouich, Lachlan Deer, Ashraf Galal Eid, Peter McGee, Daniel Schoch, Hrvoje Stojic, Jolanda Ygosse-Battisti, and Ferdinand M. Vieider. Gender effects for loss aversion: Yes, no, maybe? *Journal of Risk and Uncertainty*, 59:171–184, 2019.

C Camerer and G Loewenstein. *Behavioral economics: past, present, future.* Princeton University Press, Princeton, 2004.

T Conley and C Udry. Learning about a new technology: pineapple in Ghana. *The American Economic Review*, 100(1):35–69, March 2010.

Partha Dasgupta and Eric Maskin. Uncertainty and hyperbolic discounting. *American Economic Review*, 95(4):1290–1299, 2005.

P Dupas. Short-run subsidies and long-run adoption of new health products: evidence from a field experiment. *Econometrics*, 82(1):197–228, January 2014.

D Ellsberg. Risk, ambiguity, and the savage axioms. *The Quarterly Journal of Economics*, 75(4), November 1961.

A Foster and M Rosenzweig. Learning by doing and learning from others: human capital and technical change in agriculture. *The Journal of Political Economy*, 103(6):1176–1209, December 1995.

U Gneezy, J List, and G Wu. The uncertainty effect: when a risky prospect is valued less than its worst possible outcome. *The Quarterly Journal of Economics*, 121(4): 1283–1309, 2006.

D Kahneman and A Tversky. Prospect theory: an analysis of decision under risk. *Econometrics*, 47(2):263–292, March 1979.

G Keren and M Willemsen. Decision anomalies, experimenter assumptions, and participants' comprehension: reevaluating the uncertainty effect. *Journal of Behavioral Decision Making*, 22:301–317, 2009.

J Keynes. *A treatise on probability.* Macmillan, London, 1921.

B Kőszegi and M Rabin. A model of reference-dependent preferences. *The Quarterly Journal of Economics*, 121(4), November 2006.

M Kremer, C Null, E Miguel, and A Zwane. Trickle down: diffusion of chlorine for drinking water treatment in Kenya. *Working paper*, 2008.

Yoanna A. Kurnianingsih, Sam K. Y. Sim, Michael W. L. Chee, and O'Dhaniel A. Mullette-Gillman. Aging and loss decision making: increased risk aversion and decreased use of maximizing information, with correlated rationality and value maximization. *Frontiers in Human Neuroscience*, 9(280), 2015.

Ross Levine and Yona Rubinstein. Smart and illicit: Who becomes an entrepreneur and do they earn more? *The Quarterly Journal of Economics*, 132(2):963–1018, 2017.

Elaine M. Liu. Time to change what to sow: Risk preferences and technology adoption decisions of cotton farmers in China. *The Review of Economics and Statistics*, 95(4): 1386–1403, 2013.

D McFadden. Rationality for economists? *Journal of Risk and Uncertainty*, 19(1–3): 73–105, 1999.

M Rabin and R Thaler. Anomalies: risk aversion. *The Journal of Economic Perspectives*, 15(1):219–232, Winter 2001.

Anuj K. Shah, Eldar Shafir, and Sendhil Mullainathan. Scarcity frames value. *Psychological Science*, 26(4):402–412, 2015.

J Shogren and L Taylor. On behavioral-environmental economics. *Review of Environmental Economics and Policy*, 2(1):26–44, February 2008.

J Strauss, F Witoelar, B Sikoki, and A Wattie. The fourth wave of the Indonesian family life survey (IFLS4): Overview and field report. *WR-675/1-NIA/NICHD*, April 2009.

Mahmud Yesuf and Randall A. Bluffstone. Poverty, risk aversion, and path dependence in low-income countries: Experimental evidence from Ethiopia. *American Journal of Agricultural Economics*, 91(4):1022–1037, 2009.